MASTER'S THESIS

Welke factoren beïnvloeden de aanpak van data science projecten

Slotboom, M (Marcel)

Award date:
2020

Welke factoren beïnvloeden de aanpak van data science projecten

Which factors influence the approach to data science projects

Opleiding:	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	Marcel Slotboom
Identiteitsnummer:	
Datum:	27-06-2020
Afstudeerbegeleider	Jeroen Baijens
Meelezer	Rob Kusters
Derde beoordelaar	nvt
Versie nummer:	1.0
Status:	final

Abstract

Door de steeds verdere integratie van nieuwe technologieën, als cloud computing, sociale netwerken en mobiele technologie in de bedrijfsprocessen genereert dit een groot volume aan waardevolle data voor organisaties. Data Science geeft organisaties inzicht in de betekenis van deze data. Om deze inzichten te krijgen is een goede projectaanpak cruciaal. Als we kijken naar de toegepaste projectmethodieken in data science projecten valt het op dat deze methodieken niet of maar gedeeltelijk, vaak ad-hoc, worden toegepast door data science teams binnen organisaties. Dit resulteert in het falen van een groot deel van deze projecten. Dit onderzoek heeft een Data Science Project Framework gerealiseerd welke inzicht geeft in de project thema's die relevant zijn voor de uitvoering van data science projecten. Het framework is gevalideerd middels een case study bij een overheidsorganisatie, door het uitvoeren van expert interviews en focusgroep sessies. Dit onderzoek draagt bij aan de wetenschap door een beeld te geven van de factoren die van belang zijn voor de inrichting van de projectmethodiek voor data science projecten. Het ontwikkelde framework laat zien (1) welke projectkarakteristieken de data science projecten beïnvloeden, (2) welke thema's uit de geëvalueerde data science modellen vanuit CRSIP-DM hier een belangrijke rol in spelen om vervolgens (3) deze karakteristieken te matchen met de relevante thema's die bijdragen aan het succesvol uitvoeren van data science projecten.

Sleutelbegrippen

Data Analytics, Agile, Projectmethodieken, Projectkarakteristieken, Data science, Procesmethodieken

Samenvatting

Als we kijken naar de toegepaste projectmethodieken in data science projecten valt het op dat deze methodieken niet of maar gedeeltelijk, vaak ad-hoc, worden toegepast door data science teams. Bij organisaties welke data science projecten uitvoeren blijkt dat er vrijwel geen gebruik gemaakt wordt van een gestructureerde procesmethode voor de data analyse. Echter geven practitioners wel aan dat hun projectresultaten zullen verbeteren als ze een gestructureerde projectmethodiek gebruiken. De uitdaging van data science projecten is dan ook om een gestructureerde projectmethodiek samen te stellen op basis van de typische projecteigenschappen van data science projecten.
Om dit probleem te onderzoeken is voor deze thesis de Design Science Research (DSR) methodiek van Hevner, March, Park, and Ram (2004) toegepast als onderzoeksmethodiek. Met als doel een framework te realiseren welke inzicht geeft in de projectthema's die van belang zijn voor de diverse beïnvloedende factoren op data science projecten. Hiermee geeft het framework handvaten die bijdragen aan het project succes van data science projecten. De DSR van deze studie beantwoordt daarom de volgende onderzoeksvraag:

'Hoe beïnvloeden projectkarakteristieken de keuze voor een succesvolle projectmethodiek voor data science projecten?'

Het literatuuronderzoek uit deze studie laat zien dat er vijf categorieën zijn die invloed hebben op de manier hoe een data science project uitgevoerd wordt, dit is de project, analytische, data, organisatie en team context. CRISP-DM is hedendaags nog steeds de-facto standaard aanpak voor data science vraagstukken, echter beperkt zich dit tot de ontwikkeling en uitrol van een data science model, maar geeft dit niet inzicht in het gehele proces dat tijdens een project doorlopen moet worden. Tevens is er onderzocht in de wetenschappelijke literatuur hoe de evolutie van de methodieken die ontstaan zijn uit het CRISP-DM model omgegaan zijn met de beperkingen van het CRISP-DM model. Dit heeft geresulteerd in een vijftal relevante thema's die uit deze geëvalueerde modellen ontstaan zijn, namelijk de probleemformulering, iteratief projectmanagement, conceptualisatie, automatisering en onderhoud. Er is dus niet één standaard methodiek te hanteren, maar is afhankelijk van meerdere factoren van het data science project. Als er inzicht is in deze factoren kunnen met dit framework de belangrijke thema's aan de projectaanpak toegevoegd worden. Dit onderzoek laat zien dat voor alle type data science projecten het essentieel is om een concrete en meetbare probleemformulering vast te leggen samen met de organisatie. Scrum is een veel toegepaste agile projectmanagement methodiek binnen organisaties, maar zeker in de begin fase van het project moeten de vaste sprintduur losgelaten worden voor data science projecten en moeten flexibel gepland kunnen worden. Daarnaast is voor een routinematige project het beheer van de lifecycle van een model essentieel en draagt automatisering van het testen en de deployment bij aan de kwaliteit van het model en de snelheid van oplevering. Een conceptueel en geaccepteerd data model is belangrijk voor projecten de werken met privacy gevoelige gegevens om vroegtijdig in het project issues rondom privacy en het gebruik van gegevens te voorkomen. Voor eenmalige projecten is onderhoud, automatisering en conceptualisatie niet van belang, omdat deze vaak kort-cyclisch zijn een enkele vraag beantwoorden of een proof of concept opleveren.

Alhoewel er consensus is over de meerwaarde van dit framework tijdens de focusgroep sessies binnen de case organisatie. Zou het framework getoetst moeten worden in de praktijk door data science teams om de toepasbaarheid en volledigheid van het framework te evalueren.

Summary

Looking at the applied project methodologies in data science projects, it is striking that these methodologies are not or only partially, often ad hoc, applied by data science teams. At organizations that carry out data science projects it appears that hardly any structured process method is used for data analysis. However, practitioners do indicate that their project results will improve if they use a structured project methodology.
The challenge of data science projects is therefore to compile a structured project methodology based on the typical project properties of data science projects.
To investigate this problem, the Design Science Research (DSR) method of Hevner, March, Park, and Ram (2004) has been applied as research method for this thesis. With the aim of realizing a framework that provides insight into the project themes that are important for the various influencing factors on data science projects. In this way, the framework provides handles that contribute to the project's success of data science projects. The DSR of this study therefore answers the following research question:

"Which project characteristics influence the choice of a project methodology for data science projects?"

The literature study from this study shows that there are five categories that influence the way a data science project is executed, this is the project, analytical, data, organization and team context. CRISP-DM is still a de-facto standard approach for data science projects, but this is limited to the development and roll-out of a data science model, but does not provide insight into the entire process that must be followed during a project. The current scientific literature shows how the data science methodologies that arose from the CRISP-DM model dealt with the limitations of the CRISP-DM model. This has resulted in five relevant themes that have arisen from these evaluated models, namely problem formulation, iterative project management, conceptualization, automation and maintenance. This research shows that here is not one standard method to be used, but it depends on several factors of the data science project. If there is insight into these factors, this framework allows the important themes to be added to the project approach. This research shows that for all types of data science projects it is essential to establish a concrete and measurable problem formulation together with the organization. Scrum is a widely applied agile project management method within organizations, but certainly in the early phase of the project, the fixed sprint duration must be abandoned for data science projects and flexible planning must be possible. In addition, for a routine project, managing the lifecycle of a model is essential and automation of testing and deployment contributes to the quality of the model and the speed of delivery. A conceptual and accepted data model is important for projects working with privacy-sensitive data in order to prevent issues surrounding privacy and the use of data early in the project. Maintenance, automation and conceptualization are not important for one-off projects, as these are often short-cycle, answering a single question or providing a proof of concept.

Although there is consensus about the added value of this framework during the focus group sessions within the case organization. The framework should be tested in practice by data science teams to evaluate the applicability and completeness of the framework.

Inhoudsopgave

1. Introductie

1.1. Achtergrond

Door de steeds verdere integratie van nieuwe technologieën, als cloud computing, sociale netwerken en mobiele technologie in de bedrijfsprocessen genereert dit een groot volume aan waardevolle data voor organisaties. Deze data, ook bekend als big data (IBM, 2016; Siddiqa et al., 2016), komt met grote snelheid binnen en moet vrijwel in real-time geanalyseerd worden. Dit vereist een andere projectaanpak voor data science projecten. Bekende data science projectmethodieken, als bijvoorbeeld CRISP-DM (Chapman et al., 2000; Wirth & Hipp, 2000), welke als doel hebben om via een gestructureerd proces waarde te creëren uit de gegevens die organisaties genereren/verzamelen in hun processen, zijn in basis niet geschikt voor de verwerking van deze massale gegevens vraagstukken. Deze deels waterval-achtige projectmethodiek sluit niet aan bij deze nieuwe business behoefte, waar de vraagstelling vaak ad-hoc is en de organisatie (processen) snel resultaat verwachten. Nieuwe methodieken of aanpassing van bestaande methodieken aan de hedendaagse data science vraagstukken zijn noodzakelijk om aan de organisatiedoelstellingen te kunnen voldoen (Baijens & Helms, 2019; Gao, Koronios, & Selle, 2015; Li, Thomas, & Osei-Bryson, 2016; Saltz, 2017).

1.2. Gebiedsverkenning

Dit onderzoek vindt plaats binnen het wetenschappelijk gebied Data Science. Data science is de wetenschap van extractie van bruikbare kennis, meestal van 'Big data', dat wil zeggen veelal grote volumes van ongestructureerde of gestructureerde gegevens gegenereerd door systemen, mensen, sensoren of persoonlijke, sociale en digitale sporen van informatie van mensen (Das, Cui, Campbell, Agrawal, & Ramnath, 2015). Data science projecten hebben doorgaans als doel om correlatie en causale verbanden te ontdekken in data en om patronen en gebeurtenissen te ontdekken (Saltz, 2015).

1.3. Probleemstelling

Als we kijken naar de toegepaste projectmethodieken in data science projecten valt het op dat deze methodieken niet of maar gedeeltelijk, vaak ad-hoc, worden toegepast door data science teams. Uit recent onderzoek blijkt dat 82% van de ondervraagde respondenten geen gebruik maakt van een gestructureerde projectmethode voor de data analyse. Echter 85% van deze respondenten geeft aan dat hun project resultaten zullen verbeteren als ze een gestructureerde projectmethodiek gebruiken (Saltz, Hotz, Wild, & Stirling, 2018). Dit komt enerzijds doordat de huidige data science projectmethodieken niet iteratief genoeg zijn, maar een voornamelijk plan gedreven aanpak hebben en hierdoor niet op de problematiek van bijv. Big data kunnen inspelen (Batra, 2018; Grady, Payne, & Parker, 2017; Li et al., 2016). Anderzijds doordat men niet weet van het bestaan van een dergelijke methodiek (Saltz et al., 2018) of de methodiek sluit niet aan op het data science vraagstuk (Das et al., 2015; Gao et al., 2015; Saltz & Shamshurin, 2016). Het niet volgen van een gestructureerd proces zorgt ervoor dat er stappen vergeten worden in het analyse proces en/of worden de best-practise uit het data science werkveld niet toegepast (Saltz, 2015), hierdoor faalt een meerderheid van de data science projecten (Gartner, 2018; Saltz, 2015).

1.4. Opdrachtformulering

In de wetenschap is al veel onderzoek gedaan naar projectmethodieken en de ontwikkeling hiervan voor data science projecten (Baijens & Helms, 2019; Li et al., 2016; Mariscal, Marban, & Fernandez,

2010). In de bestaande literatuur is er nog weinig onderzoek gedaan naar de oorzaak van het niet hanteren van deze projectmethodieken en welke factoren hier invloed op hebben.
Diverse onderzoeken concluderen dat de keuze van een juiste methodologie voor een data science project sterkt afhangt van de projectkarakteristieken van een project (Ahangama & Poo, 2015b; Baijens & Helms, 2019; Gao et al., 2015; Li et al., 2016; Saltz et al., 2018).
Dit onderzoek doet een bijdrage aan de wetenschap door het beantwoorden van de volgende onderzoeksvraag:

H1 'Hoe beïnvloeden projectkarakteristieken de keuze voor een succesvolle projectmethodiek voor data science projecten?'

Om deze vraag te kunnen beantwoorden zijn de volgende deelvragen geformuleerd:

D1 'Hoe onderscheiden de hedendaagse projectmethodieken binnen het data science werkveld zich van elkaar?' (literatuuronderzoek)
D2 'Welke projectkarakteristieken hebben invloed op de aanpak van data science projecten?' (literatuuronderzoek)
D3 'Welke invloed heeft een projectmanagement methodiek op het succesvol afronden van een data science project?' (literatuur- en kwalitatief onderzoek)

1.5. Motivatie / relevantie

De traditionele data science projectmethodieken zijn veelal gebaseerd op een plan-gedreven aanpak. Echter de hedendaagse data science vraagstukken en data complexiteit vraagt om een meer iteratief karakter, met als doel snel op de veranderende omgeving (organisatie en data) te kunnen inspringen. In de bestaande literatuur is er weinig bekend over welke methodiek bruikbaar is kijkende naar de projectkarakteristieken van data science projecten (Saltz, Shamshurin, & Connors, 2017). Dit onderzoek zal een bijdrage leveren aan de huidige wetenschap door te onderzoeken welke projectkarakteristieken van een data science project invloed hebben op de keuze van de projectmethodiek. Door dit inzichtelijk te maken ontstaat er een framework, waarmee op basis van de projectkarakteristieken een passende projectmethodiek samengesteld kan worden die bijdraagt aan het succes van data science projecten. Daarnaast geeft dit onderzoek de practitioner handvaten om aan de start van het project de juiste projectmethodiek te selecteren op basis van deze projectkarakteristieken met als doel om tot succesvolle projectresultaten te komen.

1.6. Aanpak in hoofdlijnen

De Design Science Research Methodology (DSRM) van Peffers, Tuunanen, Rothenberger, and Chatterjee (2007) wordt gehanteerd als aanpak voor dit onderzoek.
Het onderzoek start met het identificeren van de probleemstelling en het formuleren van de onderzoeksvragen. Vervolgens wordt er een literatuuronderzoek uitgevoerd op basis van reeds bestaande literatuur. Het literatuuronderzoek draagt bij aan het ontstaan van een theoretisch framework dat inzicht geeft omtrent de onderwerpen data science, projectmethodieken en de contextuele eigenschappen van de data science projecten. Het literatuuronderzoek beantwoordt de deelvragen D1 en D2. De Design en Development stap uit het DSRM proces omvat de methodiek en de uitvoer van het empirisch onderzoek en moet de deelvraag D3 en de hoofdvraag beantwoorden. Tijdens de Demonstration fase worden semigestructureerde interviews en focusgroep sessies uitgevoerd om informatie te verzamelen over de huidige werkwijze en het framework te bediscussiëren. Op basis van de resultaten van het empirisch onderzoek worden er, indien

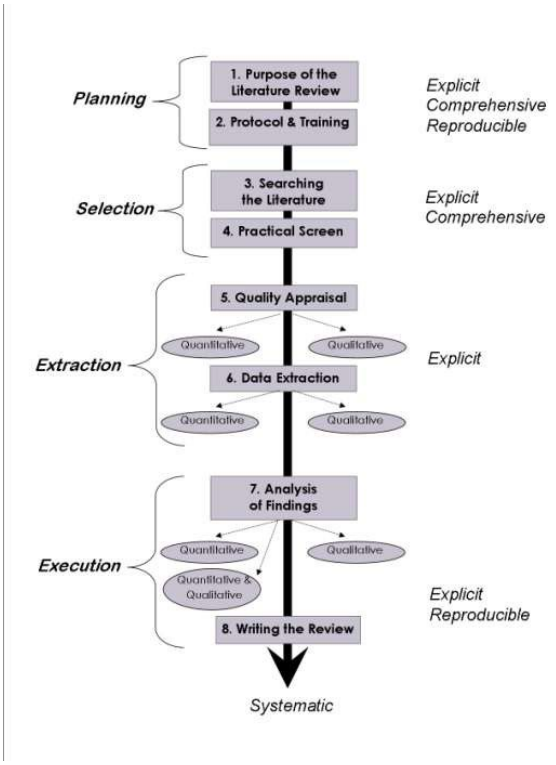
noodzakelijk, aanpassingen doorgevoerd in het framework (Evaluation). De uitwerkingen van het onderzoek worden beschreven in de thesis, welke vervolgens wordt gepresenteerd en verdedigd (Communication).

2. Theoretisch kader

Het theoretisch kader wordt gevormd door het uitvoeren van een literatuuronderzoek.
Met dit literatuuronderzoek worden de deelvragen beantwoord.

2.1. Onderzoeksaanpak

Voor het uitvoeren van het literatuuronderzoek is als leidraad de Systematic Literature Review (SLR) methodiek (figuur 2) toegepast van Okoli and Schabram (2010). In dit literatuuronderzoek zijn de acht stappen doorlopen van dit model om gestructureerd en op een wetenschappelijke wijze vorm te geven aan het literatuuronderzoek. Met als primair doel om een theoretisch kader te creëren rondom het te onderzoeken fenomeen voor het verdere verloop van het onderzoek.



Figuur 1 Systematic Literature Review (Okoli & Schabram, 2010)

Het theoretisch kader heeft als doel om inzicht te krijgen in het fenomeen data science en de bestaande projectmethodieken omtrent dit concept om vervolgens de deelvragen te kunnen beantwoorden.

2.2. Implementatie

Het van te voren vastleggen van een protocol is volgens Okoli and Schabram (2010) een cruciale stap voor het verdere literatuuronderzoek. Voor het formuleren van het protocol is de Systematic Literature Research (SLR) methodiek (Okoli & Schabram, 2010) gehanteerd. Met behulp van deze methodiek is er een protocol geformuleerd dat gestructureerd inzicht geeft in het doorlopen proces en geeft de mogelijkheid tot reproduceerbaarheid om de externe validiteit van het onderzoek te

borgen. Het protocol bevatte de volgende activiteiten:

1. Creëren van een start-set met relevante artikelen.
2. Identificatie van de zoekstrategie en de concepten.
3. Selectie van de databases.
4. Analyseren en reviewen van de resultaten.
5. Creëren van een finale set met artikelen

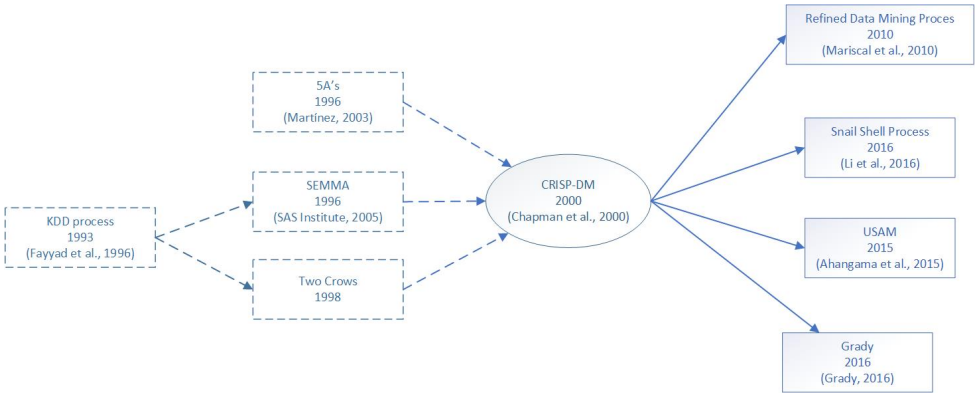
De onderzoeksbegeleiders van de OU hebben een set van zes artikelen aanbevolen om een beeld te krijgen van de concepten en terminologie in het werkveld.
Vanuit deze begrippen is een generieke query tot stand gekomen waarmee in eerste instantie gezocht is in de OU bibliotheek. Echter was er hier geen mogelijkheid om de resultaten te exporteren, daarom is er uiteindelijk voor gekozen om in de specifieke databases uit de OU bibliotheek verder te zoeken. De aanvullende literatuur is gevonden door de query uit te voeren in databases die artikelen bevatten betreffende het werkveld van Informatie Systemen en Informatica, dit waren de databases: ACM, AIS, EBSCOhost en Web of Science. De gebruikte artikelen zijn wetenschappelijk, peer-reviewed, uit een beperkte tijdframe en bevatten de begrippen die geselecteerd zijn voor het literatuuronderzoek.
Dit resulteerde in 270 artikelen, waarvan er 73 dubbel waren. Na de practical screening (op basis van title, keywords en abstract) bleven er nog 27 over. Op basis van de introductie en conclusies van de resterende artikelen bleven er uiteindelijk 13 over. Door snowballing door de 13 artikelen werden er hier nog 8 artikelen aan toegevoegd. Dit resulteerde uiteindelijk in een finale set van 21 artikelen.

In bijlage 1 van deze thesis is de gedetailleerde aanpak voor dit literatuuronderzoek beschreven.

2.3. Resultaten en conclusies

2.3.1. Projectmethodieken

Op basis van de literatuurreview betreffende data science projectmethodieken van Mariscal et al. (2010) en Baijens and Helms (2019) is er een selectie gemaakt op basis van methodieken die voortgekomen zijn vanuit CRISP-DM. De volgende modellen zijn opgenomen in dit onderzoek: Refined Data Mining Process (Mariscal et al., 2010), Snail shell process (Li et al., 2016), Ahangama and Poo (2015a) , N. Grady (2016) en Marbán, Mariscal, Menasalvas, and Segovia (2007)



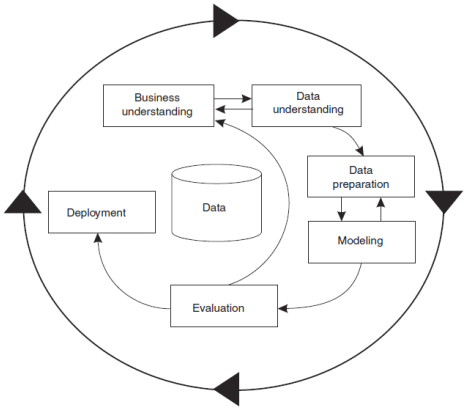
Figuur 2 Evolutie Procesmethodieken (Baijens & Helms, 2019; Mariscal et al., 2010)

De geëvalueerde methodieken voegen een aantal verbeteringen toe aan het CRISP-DM model. Eerst wordt er een korte beschrijving van het CRISP-DM model gegeven, vervolgens gaan we in op de

verschillende overeenkomstige thema's van de onderzochte modellen. Deze thema's zijn: Probleem formulering, iteratieve methodieken, conceptualisatie, onderhoud en automatisering.

CRISP-DM (Chapman et al., 2000)

Cross-Industry Standard Process for Data Mining (CRISP-DM) wordt gezien als de-facto standaard methodiek (Li et al., 2016; Mariscal et al., 2010) voor het uitvoeren van data science projecten. De belangrijkste reden hiervoor is dat CRISP-DM industrie, tool en applicatie onafhankelijk is. (Li et al., 2016; Mariscal et al., 2010; Saltz et al., 2018; Wirth & Hipp, 2000)



Figuur 3 Cross-Industry Process for Data Mining (CRISP-DM) proces model (Chapman et al., 2000)

Het CRISP-DM proces bestaat uit zes stappen (Chapman et al., 2000; Wirth & Hipp, 2000): Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment. Het proces model is een iteratief model, maar met een watervalachtig te doorlopen proces. Dit houdt in dat de fasen één voor één doorlopen worden, maar er wel beperkt iteraties naar vorige processtappen plaatsvinden.

Thema's

In de volgende paragrafen worden de onderkende thema's uit de onderzochte modellen besproken.

Iteratieve projectmethodieken

Tijdens de looptijd van projecten ontstaan er nieuwe requirements. Enerzijds door snel veranderende data en anderzijds de veranderende behoefte van de organisatie. Het is essentieel om snel op deze requirements in te kunnen spelen, zodat het de projectresultaten aansluiten bij de businessbehoefte (Ahangama & Poo, 2015a; Li et al., 2016). Een iteratieve projectmethodiek is daarom wenselijk (Mariscal et al., 2010). CRISP-DM beschrijft een waterval-achtige lifecycle voor data science projecten, maar niet de belangrijke projectmanagement activiteiten, zoals quality management en change management (Li et al., 2016). Een echte projectmethodiek is CRISP-DM ook niet, want het model laat zien hoe je het moet doen, maar niet op welke manier (Mariscal et al., 2010). Het Unified Structured Analytic model - USAM (Ahangama & Poo, 2015a) en het Snail shell process (Li et al., 2016) zijn voorbeelden van iteratieve modellen die dit onderkent hebben en beschrijven niet alleen wat je moet doen, maar ook hoe je dit moet doen. Zij voegen o.a. de fasen 'problem formulation' en 'maintenance' toe als essentiële stappen om een kwalitatief en onderhoudbaar projectresultaat te realiseren en beschrijven ook op welke manier je dit kunt toepassen.

Projectmanagement methodiek

De zojuist benoemde data science methodieken moeten toegepast kunnen worden binnen de projectmanagement methodieken. Een enkel plan-gedreven (waterval) project aanpak werkt niet voor data science projecten (Batra, 2018; Franková, Drahošová, & Balco, 2016; Li et al., 2016; Saltz, 2015). Batra (2018) argumenteert dat door gewenste innovatieve en iteratieve manier van ontwikkelen, de agile projectaspecten belangrijker zijn dan de plan-gedreven aspecten. Binnen data science projecten is een enkel watervalmethode, zeldzaam of zelfs afwezig. Batra (2018) onderkent twee significante projectmanagement varianten: De agile-heavy en de agile-plan balanced. De agile heavy is een methodiek die letterlijk het Agile manifest (Beck et al., 2001) volgt, waar individualiteit, interactie en zelfgeorganiseerde aanpak belangrijke kenmerken zijn. De agile-plan balanced is een deels planmatige aanpak waarbij agile principes toegepast worden, dus een hybride methodiek die toepasbaar is voor risicovolle projecten die vaak business case gedreven zijn. Hierbij past een iteratief proces waar continue de business bij betrokken is, maar er wel een duidelijke planmatige projectsturing nodig is om de projectgrenzen te bewaken. Agile principes in de agile-plan balanced variant zie je voornamelijk terug in de bruikbare onderdelen van agile methodieken als Scrum en KanBan (Batra, 2018). De Agile-heavy variant worden voornamelijk toegepast in organisaties waar geen directe sturing is van top management, vaak kleine organisaties/teams waar men veel vrijheid heeft en het afbreukrisico voor het project relatief laag is. Tevens zie je dit bij projecten die kort cyclisch zijn en vaak innovatief gedreven zijn bijv. met enkel als doel het realiseren van proof of concepts. Deze projecten zitten niet te wachten op langlopende en gestructureerde processen en zijn vaak ook eenmalig van karakter (Batra, 2018; Saltz, 2015).

Scrum en Kanban

Scrum is een agile framework voor het 'ontwikkelen, leveren en onderhouden van complexe producten' (Schwaber & Sutherland, 2017). Het verdeelt een groot complex project in een serie van kleine projecten, welke men 'sprints' noemt. Sprint zijn in Scrum vastgestelde tijdblokken meestal tussen de 2 en 4 weken in lengte waarin de kleine projecten opgeleverd gaan worden. Gedurende een sprint communiceert het team intensief met elkaar en houden daily stand-ups. Dit zijn korte meetings waarin ieder teamlid aangeeft wat zijn voortgang is en of er problemen zijn (impediments). Aan het eind van elke sprint demonstreert het team de resultaten aan de stakeholders en verzamelt feedback (sprint review). Om een sprint af te sluiten reflecteert het team de afgelopen sprint en bekijkt of er verbeteringen noodzakelijk zijn voor de volgende sprint (sprint retrospective).
Kanban heeft daarentegen geen strak gedefinieerd proces en benoemd ook geen projectrollen. Kanban heeft als doel om de werkstromen te visualiseren en het beperken van 'work-in-progress' (Anderson, 2010). Het visualiseren wordt gedaan door een Kanban bord, hierop zijn de diverse fasen te zien van het werk. Meestal zie je hier kolommen (lanes) als 'To Do', 'Doing' en 'Done'. De werkopdrachten worden dan bijvoorbeeld met geeltjes in de betreffende kolom geplakt. Er is dan altijd een visuele weergave voor het projectteam over het werk onderhanden en de status hiervan.
Ten slot zijn er methodes die Scrum en KanBan combineren. Dit doen ze op de volgende manieren; ten eerste door de duur van een iteratie aan te passen aan de project fase. Als tweede ontkoppel de overlegstructuur los van de iteraties en als derde maak alleen een globale planning voor de project items (Saltz & Suthrland, 2019).

Data science methodieken

Li et al. (2016) en Mariscal et al. (2010) hebben in het model de **probleem formulering** als een cruciale stap voor het slagen van het data science project toegevoegd. Deze stap realiseert een duidelijke, afgestemde, concrete en meetbare probleemdefinitie, waarmee een duidelijk toetsing (evaluatie) van het resultaat kan plaatsvinden gedurende het project.

Ahangama and Poo (2015a) introduceren het begrip **conceptualisatie** in hun USAM methodiek. Doordat dit model primair is ontwikkeld voor de gezondheidszorg, waar privacy en security belangrijk requirements zijn, hanteert dit model een conceptueel data model dat een afspiegeling geeft van de werkelijkheid. Hierdoor kan men eerst beoordelen of de informatie compleet en relevant is alvorens de beoogde verwerking van de gegevens plaatsvindt. Dit heeft als doel om te voorkomen dat de modellen gecreëerd uit deze data beslissingen nemen op basis van ongewenst gebruik van bijzondere persoonsgegevens en voorkomt issues als bijvoorbeeld etnische profilering. Deze conceptualisatie is ook toepasbaar in werkgebieden waar privacy van belang is als bijvoorbeeld overheidsorganisaties of verzekeraars.
Mariscal et al. (2010) introduceren de processtap **automate** (automatiseren) in hun model. Dit is gebaseerd op het idee om gebruikers zonder data science expertise eerder verkregen modellen toe te laten passen op nieuwe data. Ook is het automatiseren van je testprocedures een belangrijke stap om tot een betere kwaliteit van het product te komen (Das et al., 2015). Als je kijkt naar product georiënteerde projecten worden deze veelal automatisch uitgerolt in primaire processen van organisaties.
Omdat bedrijfsprocessen afhankelijk zijn van deze services moeten deze data science producten **onderhouden** worden. Onderhoud van een model of dashboard is essentieel voor de bedrijfsprocessen, omdat omstandigheden continue kunnen wijzigen, dit kan voortkomen uit een wijzigende behoefte vanuit de organisatie of de aanpassing van een model door wijzigende wet- en regelgeving. Deze processtap beheert de lifecycle van het data science model: Model selectie, gebruik, wijzigingen, vervanging en uitfasering (N. Grady, 2016; Li et al., 2016), maar tevens beheertaken als data backups, data mining model updates en software updates (Mariscal et al., 2010).

Resultaat onderzochte data science methodieken

Tabel 1 betreft een overzicht van alle onderzocht methodieken en zijn de opvallende verschillen tussen de modellen inzichtelijk gemaakt (grijze arcering).

CRISP- DM	(Mariscal et al., 2016)	(Ahangama et al. ,2015)	(Li et al., 2016)	(Grady, 2016)
	Life Cycle Selection		Business Understanding	
Business understanding	Domain Knowledge Elicitation	Project initiation	Business Understanding	Plan
	Human Resource Identification	Domain understanding		
	Problem Specification		Problem formulation	
Data Understanding	Data Prospecting	Data Understanding	Data Understanding	Collect
		Conceptualization		
Data preparation	Data cleaning	Data Preparation	Data Preparation	Curate
	Preprocessing			
	Data reduction and projection			
Modeling	Choosing the DM Task	Data modeling	Modeling	
	Choosing the DM Algorithm			
	Build model			
	Improve model			
Evaluation	Evaluation	Validation	Evaluation	Act
	Interpretation			
Deployment	Deployment	Presentation	Deployment	
	Automate			
	Establish On-Going support	Presentation	Maintenance	Act

Tabel 1 Overzicht van toevoegingen model t.o.v. CRISP-DM

We kunnen op basis van de literatuur de volgende projectmethodiek thema's vastleggen:

- *Problem Formulation*: Heldere en concrete weergave van het business probleem
- *Iteratieve methodieken* binnen data science en projectmanagement methodieken.
    - Agile-heavy, een projectmethodiek waar er vrijwel tussen alle fasen iteraties mogelijk zijn en snel op veranderende omstandigheden kunnen inspelen.
    - Agile-plan balanced, een deels planmatige aanpak, gebruikmakende van toepasbare agile principes.
- *Conceptualization:* Acceptatie van het datamodel alvorens verder te gaan met de modeling fase.
- *Maintenance*: Complete lifecycle van de modellen (monitoren, onderhoud, vernieuwen en uitfaseren)
- *Automate*: Hoge mate van geautomatiseerde model implementaties (voorbrengingsproces en testcyclus).

2.3.2. Projectkarakteristieken

Een project karakteristiek is een specifieke eigenschap van een project, bijvoorbeeld het type data of de hypothese vorm is een karakteristiek welke het verloop van een project kan beïnvloeden (Saltz et al., 2017).

Projectkarakteristieken kunnen we in vijf contexten onderverdelen: Project context, data context, analytics context, team context en de organisatie context (Saltz & Shamshurin, 2015; Saltz et al., 2017).

Project Context

Saltz and Shamshurin (2015) definiëren twee type data science projecten, namelijk routinematige en eenmalige projecten. Het doel van een routinematig project is het creëren van een product om het bedrijfsproces te ondersteunen, waarbij onderhoud en beheer ook belangrijke elementen zijn van de project lifecycle (Ahangama & Poo, 2015a; Li et al., 2016; Mariscal et al., 2010).

Bij eenmalige projecten is het doel vaak innovatie of ontstaan uit een adhoc vraag of probleem vanuit de organisatie die snel beantwoord moet worden. De uitkomst is van te voren moeilijk te voorspellen en hebben vaak als doel om nieuwe kennis te verkrijgen, daarnaast zijn deze projecten moeilijk te plannen in een gestructureerd proces en zijn daarom vaak ook kort cyclisch van aard en minder gestructureerd dan de routinematige projecten (Saltz & Shamshurin, 2015). Het afbreukrisico is bij een innovatieve project een stuk kleiner. Een eenmalig resultaat of een proof of concept is vaak het resultaat van een dergelijk project.

Analytics context

De gebruikte projectmethodiek is afhankelijk van het data science vraagstuk. Primair kunnen we hier twee type vraagstukken onderkennen: Hypothese testing en hypothese generation (Das et al., 2015; Saltz, 2015; Saltz et al., 2017). Bij Hypothese testing is het probleem reeds gedefinieerd, men weet waar men naar op zoek is in de data. Hypothese testing projecten hebben veelal gestructureerde data als input. Hierbij worden vaak supervised methodieken als bijvoorbeeld classificatie en regressie toegepast. De hierbij ontwikkelde modellen worden vaak in bedrijfsprocessen gehanteerd en als een service aangeboden.
Daarentegen zijn hypothese generating vraagstukken exploratief van aard en weet men niet van te voren wat men in de dat tegenkomt of wat het doel is, hier worden veelal unsupervised methodieken als clustering en profiling toegepast op datasets (Provost & Fawcett, 2013; Richarz et al., 2019). Het doel van hypothese generating projecten zijn voornamelijk het vinden van onbekende patronen of relaties in data. In deze projecten worden meer agile achtige projectmanagement methodieken gehanteerd en is de return of investment (ROI) moeilijk te voorspellen. Deze projecten zijn vaak kleinschaliger of korter in duur en hebben hierdoor een kleiner afbreukrisico (Saltz et al., 2017).

Data context

Big en small data zijn termen die je sinds 2008 tegenkomt (Kitchin & Lauriault, 2015; Saltz et al., 2017). Alle data wordt standaard gezien als small data ongeacht het volume. De term ‘big’ is daarom soms misleidend en is gekarakteriseerd als veel meer dan volume. Sommige ‘small’ datasets kunnen zeer groot in omvang zijn. Daarentegen is small data vaak gelimiteerd in volume en snelheid, maar heeft een lange geschiedenis van ontwikkeling uit bijvoorbeeld bedrijfsprocessen of bekende bronnen en is hierdoor veel meer geschikt om specifieke onderzoeksvragen te beantwoorden (Kitchin & Lauriault, 2015; Saltz et al., 2017). De afgelopen jaren is het volume van data wereldwijd enorm toegenomen door het continue genereren van enorme hoeveelheden van heterogene, gestructureerde en ongestructureerde data, de zogenaamde ‘Big Data’ (Siddiqa et al., 2016). Dit vergt een andere project aanpak, daarvoor moeten we kijken naar de eigenschappen van deze data. Big data kenmerkt zich door de enorme volumes (terabytes of petabytes); de hoge snelheid, gecreëerd in of bijna real-time; diversiteit in variëteit, gestructureerd en ongestructureerd, en vaak tijdelijk van aard (Batra, 2018; Gao et al., 2015; Kitchin & Lauriault, 2015). Deze big data

vraagstukken eisen een iteratief proces, waar snel op veranderingen ingespeeld kan worden (Li et al., 2016; Mariscal et al., 2010).
Privacy kaders vereisen dat gegevens alleen verwerkt mogen worden als de doelmatigheid aangetoond kan worden (EU, 2016). In het analyse proces moet hier dus tijdens de data verzameling rekening mee gehouden worden om privacy en ethische problematiek te voorkomen, conceptualisatie en automatisering van processen dragen hieraan bij (Ahangama & Poo, 2015a; Grady et al., 2017; Mariscal et al., 2010).

Team context

Voor data science teams is de samenstelling van het team essentieel (Franková et al., 2016; Gao et al., 2015; Li et al., 2016). Het moet een multidisciplinair team zijn waar zowel de business als de technische, analytische en/of statistische expertise aanwezig zijn, afhankelijk van het data science vraagstuk (Bach, Zoroja, & Celjo, 2017; Gao et al., 2015). Daarnaast is een transparante en open communicatie binnen een team essentieel, hiermee onderstaat er een gedeelde kennis en kan snel op wijzigende omstandigheden ingespeeld worden (Asadi Someh, Breidbach, Davern, & Shanks, 2016). Een voorbeeld van open communicatie is het toepassen van bepaalde aspecten uit KanBan of Scrum als het KanBan bord of de daily stand-ups uit Scrum, maar ook de toepassing van een multidisciplinair team draagt hier aan bij, hierdoor ontstaan korte communicatie lijnen.

Organizational context

Saltz et al. (2017) stelt dat een organisatie met een ROI (Return of Investments) gefocuste cultuur, waar dus een valide business case noodzakelijk is voor de start van een project, impact heeft op het project. Dit heeft in bijzonder impact op projecten met een hypothese generating context, deze zijn moeilijk van de grond te krijgen, doordat de resultaten moeilijk in te schatten zijn. Een ROI gedreven organisatie waar het top management veel invloed heeft zal eerder een hybride project methodiek hanteren. Daarnaast zal een organisatie waar deze top management invloed niet aanwezig is vaker met agile methodieken werken (Batra, 2017). Ook de organisatie omvang heeft invloed op de manier van werken. Kleine organisaties zullen eerder op agile principes terugvallen om tijdig tot resultaten te komen, waarbij het financiële risico vaak kleiner is dan grote organbisaties met complexe data science vraagstukken en een veelvoud aan bronnen (Batra, 2017; Saltz et al., 2017).

Conclusie

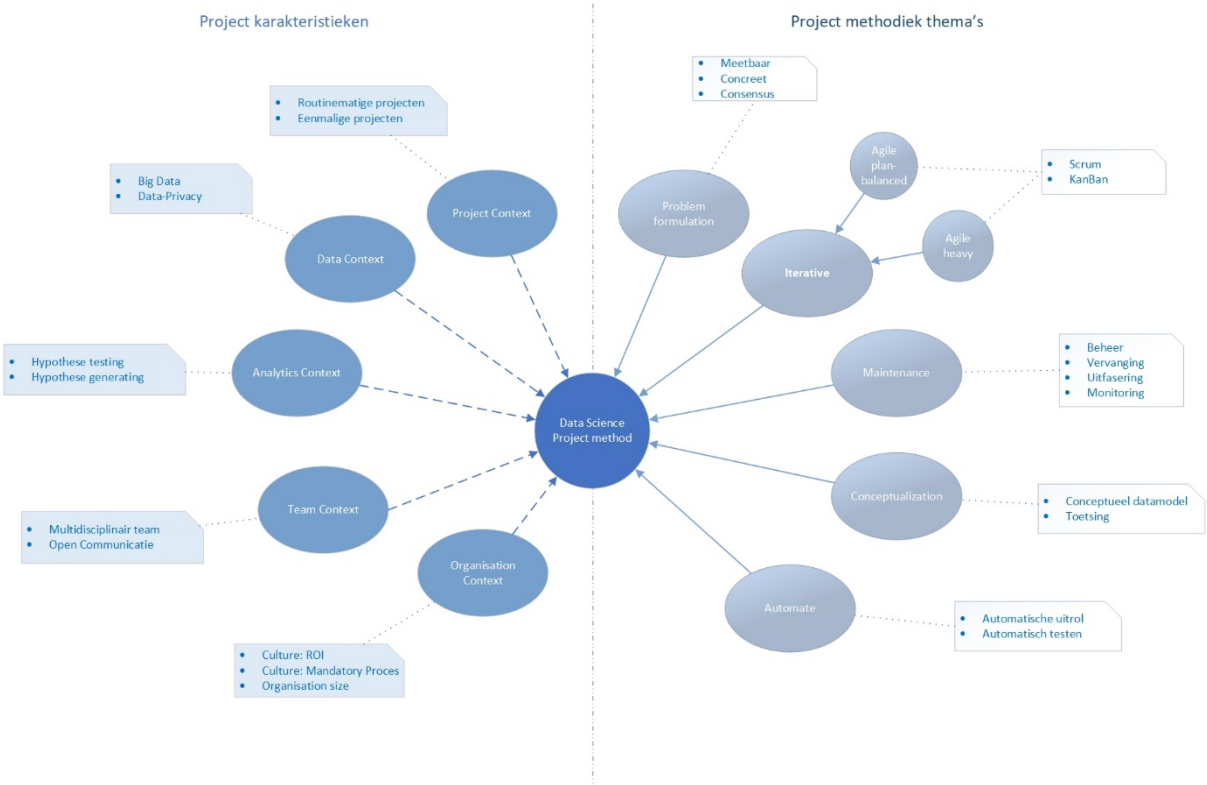
Deze paragraaf brengt de resultaten uit het literatuuronderzoek samen, om vervolgens dit verder uit te werken naar het artefact. Onderstaande tabel geeft de contextuele eigenschappen weer die invloed hebben op data science vraagstukken en invloed hebben op het succes van het project.

Categorie	Karakteristiek
Project Context	Routinematige projecten Eenmalige projecten
Analytical context	Hypothese testing Hypothese generating
Data context	Big Data Data Privacy
Team context	Multi-disciplineer team Open communicatie
Organizational context	Grote van de organisatie (groot, klein) Organisatie cultuur; ROI Organisatiecultuur; Proces

Tabel 2 kritische project karakteristieken per aandachtsgebied

2.3.3. Data science project methodologie framework

Op basis van het literatuuronderzoek kunnen we concluderen dat er twee categorieën invloed hebben op de uitvoer van projecten. De projectkarakteristieken, oftewel de projectcontext en anderzijds de projectmethodiek thema's. In het theoretisch framework zijn deze project beïnvloedende factoren samengevoegd tot één referentiemodel, het zogenaamde Artefact (figuur 4). Dit model wordt in het empirisch onderzoek toegepast als referentie model en zal geëvalueerd worden tijdens de case study.



Figuur 4 Theoretisch model – Data Science Project Methodologie Framework

2.3.4. Doel van het vervolgonderzoek

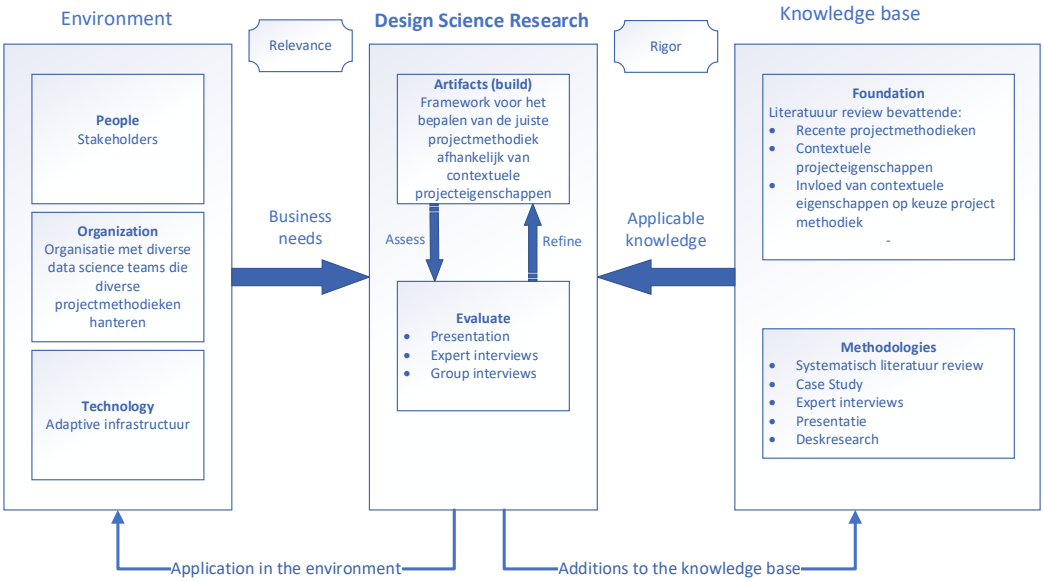
Door het ontbreken van een duidelijk framework voor de selectie van een data science projectmethodiek op basis van de projectkarakteristieken, maken organisaties verkeerde keuzes in projecten of hebben geen weet van de best-practises in dit werkveld. Veelal met als gevolg vertraagde of falende projecten. De practitioners hebben behoefte aan een framework die richtlijnen geeft in de selectie van de methodiek die het best passend is voor het project.
Daarom is het doel van dit Design Science onderzoek om 1) een framework te realiseren dat data science projectteams richtlijnen aanbiedt voor de juiste selectie van meest passende projectaanpak, 2) inzicht geeft in beïnvloedende projectkarakteristieken rondom data science projecten en 3) door de implementatie van de relevante thema's het project tot een succes brengt.

3. Methodology

In dit hoofdstuk zijn de methodieken die toegepast zijn binnen dit onderzoek beschreven en verantwoord. Het onderzoek is een design science research (DSR) waar de Design Science Research Methodology (DSRM) van Peffers et al. (2007) is toegepast. Om de wetenschappelijke aanpak en de relevantie met de omgeving te waarborgen is het Design Science Framework van Hevner et al. (2004) toegepast.

3.1. Conceptueel ontwerp: keuze van onderzoeksmethode(n)

Dit onderzoek realiseert een framework welke toepasbaar is voor practitioners en bijdraagt aan de kennis in de wetenschap. Design Science Research sluit hierop aan, door op basis van het geformuleerde theoretisch framework (het artefact) een evaluatie uit te voeren in de natuurlijke omgeving waar dit fenomeen optreedt d.m.v. een case study. Design Science richt het onderzoek op het bouwen en evalueren van artefacten welke ontwikkeld zijn voor de gedefinieerde business behoefte (Hevner et al., 2004; Peffers et al., 2007). Hevner et al. (2004) hebben een framework ontwikkeld dat bijdraagt aan het inzichtelijk maken van de performance van design-science onderzoek in Information Systems door richtlijnen te formuleren voor het begrijpen, uitvoeren en evalueren van het onderzoek. In figuur 5 is dit onderzoek geplot op het Design Science framework.

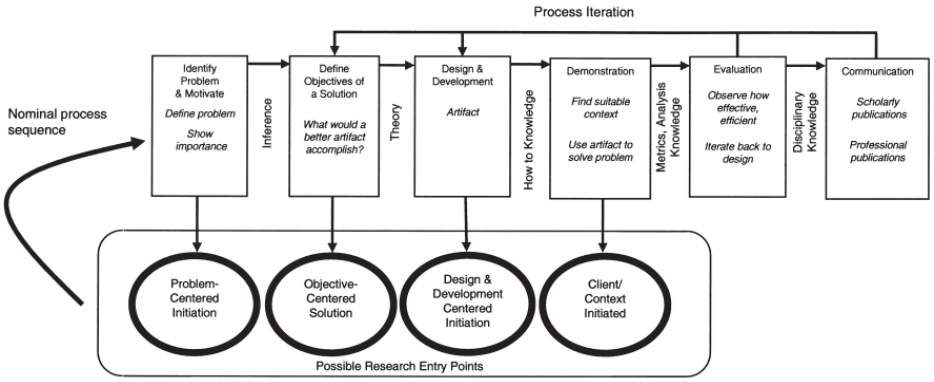


Figuur 5 Design Science Framework (Hevner et.al. 2004)

De wetenschappelijke bijdrage die benodigd is om het artefact te creëren, de zogenaamde knowledge base, is gebaseerd op een basis (foundation) van de beschikbare wetenschappelijke literatuur van het te onderzoeken onderwerp. De evaluatie van het model wordt getoetst door het uitvoeren van een Case Study in een organisatie.

3.2. Technisch ontwerp: uitwerking van de methode

Pfeffer et.al. (2007) hebben een framework ontwikkel voor de productie en presentatie van een Design Science onderzoek voor Informatiesystemen, de zogenaamde Design Science Research Methodology (DSRM). Dit framework helpt onderzoekers hun onderzoek te presenteren op een gestructureerd en begrijpbaar referentie framework, het artifact. Het procesmodel bestaat uit 7 stappen: Identify Problem & Motivate, Define objectives of a solution, Design & Development, Demonstration, Evaluation en Communication.
In onderstaand figuur is het procesmodel weergegeven.



Figuur 6 DSRM Process Model (Peffers et al., 2007)

Voor de aanpak van deze Thesis zijn de processtappen uit het procesmodel doorlopen. In tabel 3 is een globaal overzicht van de Thesis structuur geplot op de processtappen uit de design science methodiek. In de volgende paragrafen worden de processtappen beschreven zoals deze wordt gehanteerd in dit onderzoek.

3.2.1. Identify Problem, motivation and objectives

Het eerste hoofdstuk uit deze thesis definieert de onderzoeksvraag en deelvragen welke vanuit de probleemstelling is geformuleerd. De doelstelling van het onderzoek komt voort uit de probleemstelling en het theoretisch kader, welke de mogelijkheid en haalbaarheid weergeven van het onderzoek. Het onderzoek start met een literatuuronderzoek. In het literatuuronderzoek is de relevante context en inhoud beschreven en op basis van deze informatie is het theoretisch kader, het artefact gerealiseerd. Het theoretisch kader beantwoord de deelvragen van het onderzoek. In het empirisch onderzoek zal dit theoretisch kader bijdragen aan het beantwoorden van de hoofdvraag.

3.2.2. Design & Development

Om het framework te ontwikkelen is een literatuuronderzoek uitgevoerd om zo een ‘knowledge base’ te creëren van reeds uitgevoerd wetenschappelijk onderzoek omtrent het onderwerp ‘data science projecten’. De doelstelling van dit onderzoek is om een framework te realiseren dat de impact van bepaalde beïnvloedende projectkarakteristieken op data science projecten te vertalen in thema’s die relevant zijn voor het succesvol slagen van het project.

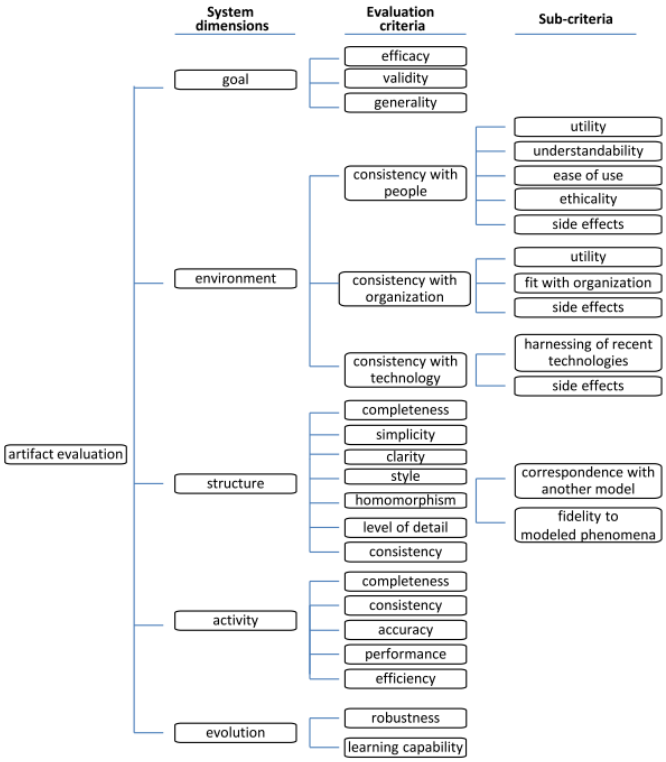
3.2.3. Demonstration

Een demonstration laat zien of het gebruik van het artefact één of meerdere problemen oplost (Peffers et al., 2007). De waarde van het artefact gecreëerd uit het theoretisch kader zal gedemonstreerd worden op basis van de resultaten van de case study. Het doel van de demonstration is het verzamelen van data voor het beantwoorden van de onderzoeksvraag. De dataverzameling zal plaats vinden door het uitvoeren van twee iteraties. De eerste iteratie betreft semigestructureerde interviews met experts uit diverse data science teams. In deze interviews wordt data verzamelen over de gehanteerde werkwijze en aanpak van de data science vraagstukken door de betreffende data science teams. Deze iteratie zal plaatsvinden met medewerkers die een overview hebben over de gehele projectcyclus van de data science teams van het betreffende bedrijfsonderdeel waar zij toe behoren. De input van deze iteratie zal verwerkt worden en dient als input voor de tweede iteratie, de focusgroep interviews. In de tweede iteratie zullen de diverse thema's van het framework in relatie tot de projectkarakteristieken getoetst worden op relevantie in de praktijk.

3.2.4. Evaluation

Voor de evaluatie wordt het FEDS (Framework for Evalution in Design Science) framework van Venable, Pries-Heje, and Baskerville (2016) toegepast. Het onderzoek start met een artificieel evaluatie vanuit de wetenschappelijke literatuur (the rigor). Echter willen wij de resultaten uit de literatuur (het artefact) toetsen op de praktische relevantie, dit doen wij met een naturalistische evaluatie, namelijk de Case study. De evaluatie zal dus sociaal en user-centric georiënteerd zijn, kijkende naar het FEDS framework dus een 'Human Risk & Effectiveness' evaluatiestrategie. De evaluatie vindt plaats door het uitvoeren van een formatieve en summatieve evaluatie. Een formatieve evaluatie draagt bij aan het verbeteren van het design proces. Deze is uitgevoerd door een documentonderzoek en het uitvoeren van expert interviews en de focusgroepsessies, welke als input dienen voor de summatieve evaluatie, de resultaten zijn weergegeven in 6.1. De summatieve evaluatie heeft als doel om het framework middels focusgroep sessies te valideren op o.a. de accuraatheid, helderheid en toepasbaarheid in de praktijk.

Voor de summatieve evaluatie is het framework van Prat, Comyn-Wattiau, and Akoka (2014) gehanteerd om een kwalitatieve evaluatie uit te voeren. De studie van Prat et al. (2014)introduceert een holistische view van de evaluatie criteria van een Design Science artefact. Deze criteria zijn onderverdeeld in een vijftal dimensies: doel, omgeving, structuur, activiteit en evolutie (zie figuur 7). Op basis van de onderzoeksresultaten worden de relevante dimensies en criteria die van toepassing zijn op dit DSR onderzoek geselecteerd. De uitwerking van de summatieve evaluatie staat in 6.3.



Figuur 7 Evaluatie dimensies en criteria voor artefact evaluatie (Prat et al., 2014).

3.2.5. Communication

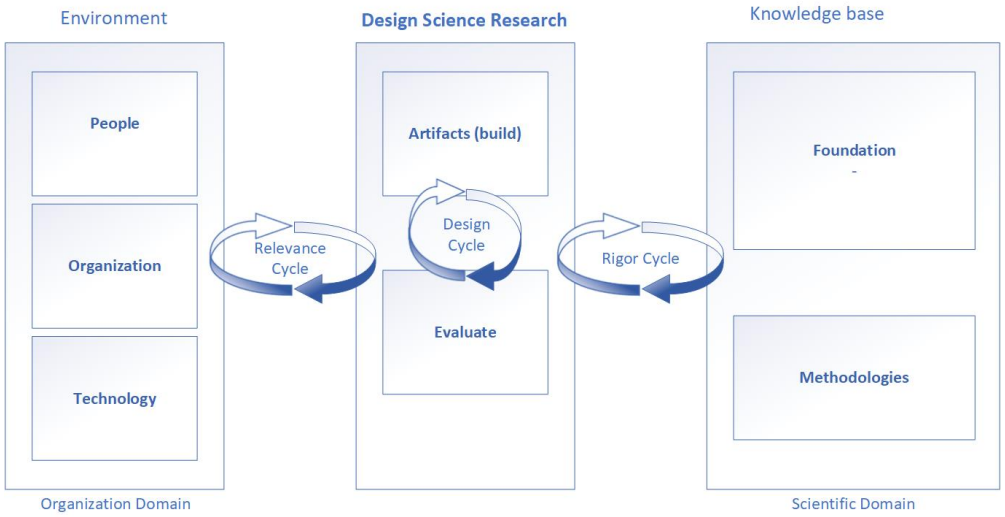
De laatste stap van het onderzoek is de communicatie van het probleem, de relevantie, het artefact zijn toepasbaarheid en wat het toevoegt aan de body of knowledge en voor de practisioners.
Dit wordt gerealiseerd door de deliverables van dit onderzoek d.m.v deze thesis. Onderstaande tabel geeft een overzicht van de Design Science Research zoals deze is toegepast in deze thesis:

Proces stap	Input	Relatie met de Thesis	Output
Identify Problem & Motivate	Literatuur (voor)onderzoek	H1 Onderzoek motivatie en onderzoeksvragen	Probleembeschrijving
Define Objectives of a Solution	Probleembeschrijving	H2 Literatuuronderzoek, Theoretisch kader	Doelstellingen Theoretisch framework (artefact)
Design & Development	Theoretisch framework Survey Expert Interviews	H3 Methode van onderzoek H4 Design	Interview guide Concept van het framework voor het empirische onderzoek
Demonstration	Framework (concept) Presentatie (validatie)	H5 Demonstration	Data verzameling Iteraties
Evaluation	Feedback presentatie Lessons learned	H6 Evaluatie H7 Conclusies en aanbevelingen	Aangepast framework door diverse iteraties
Communication	Thesis verslag Thesis verdediging	Thesis en eindpresentatie	Thesis cijfer

Tabel 3 Onderzoeksproces en de resultaten

3.3. Reflectie t.a.v. relevance and rigor

Terwijl traditionele onderzoeksbenaderingen zich richten op de betrouwbaarheid en validiteit van bevestigde hypothesen, concentreert Design Science Research (DSR) zich op een wetenschappelijk onderbouwt artefact. Dit artefact wordt geëvalueerd op kwaliteit, effectiviteit en nut via goed uitgevoerde evaluatiemethodieken. Hevner et al. (2004) beschrijft DSR als een iteratief en incrementeel probleemoplossend proces dat rigor (wetenschappelijke domein) en relevance (organisatie domein) met elkaar verbindt.



Figuur 8 Design Science Cycles (Hevner, 2007)

De Rigor cycle voegt bestaande wetenschappelijke kennis van het domein toe aan het DSR project om te borgen dat het ontwikkelde artefact een bijdrage levert aan het wetenschappelijk domein. De Relevance cycle borgt de eisen van het organisatie domein en bepaalt de acceptatie criteria voor de evaluatie van de onderzoeksresultaten. De basis van het onderzoek ligt in de Design cycle hier wordt het artefact gecreëerd en geëvalueerd op basis van meerdere iteraties tussen build en evaluatie. Figuur 8 toont de drie design science cyclussen van Hevner (2007) geplot op het Design Science Framework (figuur 5). Om de rigor te borgen van het in dit onderzoek ontworpen artefact, zal het DSR project als uitgangspunt het artefact hanteren en deze toetsen in het organisatie domein d.m.v. expert interviews en een focusgroep interview tijdens de case study. Tijdens de evaluatie van de resultaten uit de interviews zal het artefact aangepast worden om de synergie tussen het organisatie en wetenschappelijk domein te borgen.

4. Design

4.1. Inleiding

In dit hoofdstuk bespreken we het ontwerp van het framework dat is samengesteld vanuit de resultaten van het literatuuronderzoek. In dit framework worden de projectkarakteristieken uitgezet tegenover de thema's die geconstateerd zijn in de onderzochte modellen. Deze methodiek maakt het mogelijk om op basis van projectkarakteristieken inzicht te krijgen in de relevante thema's voor het data science vraagstuk. Deze thema's kunnen vervolgens aan de projectaanpak worden toegevoegd.

4.2. Thematisch projectkarakteristiek framework

De verschillende projectkarakteristieken hebben allen invloed op de manier hoe data science projecten worden aangepakt. Per karakteristiek (context), project, analytisch, data, organisatie en team zal beschreven worden wat de impact is op de onderkende thema's uit de onderzochte methodieken.

4.2.1. Project context

Routinematige projecten worden veelal ontwikkeld voor integratie in bedrijfsprocessen. In deze bedrijfsprocessen is over het algemeen weinig data science kennis aanwezig, daarom is een geautomatiseerde implementatie in deze processen noodzakelijk. Door trends in de samenleving en/of de organisatie veranderen modellen continue, daarom moeten de modellen onderhouden worden. Hiervoor moeten processen ingericht worden rondom de lifecycle van deze modellen. Naast routinematige projecten zijn er ook projecten die moeten voorzien in nieuwe inzichten en kennis van een vaak adhoc fenomeen, de eenmalige projecten. Dit zijn projecten die moeilijk in tijd te plannen zijn en waar het resultaat lastig van de voren te voorspellen is. Hierdoor is het complex om een valide business case te verantwoorden voor dit type projecten, daardoor is er een groot afbreukrisico aanwezig (Saltz et al., 2017). Hierbij past een agile projectmethodiek waar snel op veranderingen ingespeeld kan worden, daarnaast omdat dit vaak eenmalige vraagstukken zijn is onderhoud en automatisering minder van belang, maar meer het resultaat van de analyse. Kanban is meer effectief, dan Scrum voor data science projecten als deze toegepast worden in exploratieve onderzoeken (Saltz et al., 2018). Bij een meer sprint gebaseerde aanpak als Scrum is voor deze projecten moeilijk in te schatten hoe lang de realisatie van een taak duurt, dit in tegenstelling tot de routinematige projecten.

4.2.2. Analytische context

Bij een hypothese testing vraagstuk is het doel van de analyse reeds bekend, men weet welke informatie men uit de data wil halen. Vaak zijn dit repeterende processen waar de aanpak bekend is (Saltz et al., 2017), maar waar het wel mogelijk moet zijn om wijzigende omstandigheden tijdens het project op te kunnen vangen (Li et al., 2016). Een iteratieve projectaanpak is hier toepasselijk. Een hybride projectaanpak past hierbij om enerzijds plan-gedreven te kunnen werken maar waar het wel mogelijk is om binnen de diverse projectstappen met agile principes te kunnen werken (Batra, 2018). Bij hypothese generating projecten waarbij het onduidelijk is wat de analyse zal opleveren past een meer agile aanpak (Saltz et al., 2017). Deze projecten zijn heel moeilijk in tijd te plannen, doordat men van te voren niet weet welk data men nodig heeft en welke modelering hierbij past (Saltz & Suthrland, 2019).

4.2.3. Data context

Big Data projecten hebben te maken met grote hoeveelheden data die in een zeer korte periode verwerkt moeten worden. Dit vraagt om een zeer iteratieve aanpak, waarbij snel op de veranderende data ingesprongen kan worden. Voor big data projecten is een agile aanpak essentieel, zodat snel op de veranderende data geanticipeerd kan worden tijdens het project (Gao et al., 2015; Saltz, 2015; Saltz & Shamshurin, 2016). Tevens is het belangrijk om een hybride projectaanpak te hanteren voor risicovolle big data projecten, waar agile methodieken in een geplande aanpak samenkomen. Bij small data projecten waar de data beter inzichtelijk is, is een meer gestructureerde projectaanpak wenselijk, maar wel met een zodanige iteratieve aanpak dat men makkelijker op veranderingen kan inspelen. Hierbij past tevens een hybride projectaanpak waar

op basis van een gestructureerde aanpak agile methodieken gehanteerd kunnen worden, denk hierbij o.a. aan daily stand-ups, sprints en hulpmiddelen als Kanban borden (Gao et al., 2015; Saltz, 2015; Saltz & Shamshurin, 2016). Kijkende naar privacy en ethiek, is een gestructureerde aanpak wenselijk, zodat tijdens de data preparatie al consensus is van alle stakeholder over de gegevens welke gebruikt mogen, een conceptueel datamodel voor de daadwerkelijke modelering kan hierbij ondersteunen.

4.2.4. Organisatie context

In business case gedreven organisaties is het de doelstelling van het management om het afbreukrisico van een project te minimaliseren. Er wordt daarom door top management gestuurd op een gestructureerde en beheersbare projectaanpak. Daarnaast is het juist vanwege het afbreukrisico belangrijk dat snel op veranderingen en nieuwe inzichten ingespeeld kan worden tijdens het project. Hiervoor is het noodzakelijk om naast een planmatige aanpak ook agile principes te hanteren. Je ziet dan vaak dat de data verzameling planmatig is en het modeleren een agile aanpak heeft (Batra, 2018). Daarnaast speelt ook de grote van de organisatie een rol. In kleine organisaties zie je dat agile-methodieken frequenter toegepast worden, mede doordat teams kleiner zijn, er minder management sturing is en het afbreukrisico vaak kleiner is. In grote organisaties zijn teams vaak multidisciplinair van samenstelling, zijn er specialistische teams die specifieke taken uitvoeren waar zowel de data scientist, als organisatie specialisten in één team zitten (Gao et al., 2015; Saltz et al., 2017).

4.2.5. Team context

Communicatie is belangrijk in projecten. In grote organisaties kunnen diverse taken in het proces door verschillende teams uitgevoerd worden. Denk hierbij aan de aanlevering van data, het modeleren van de data, het implementeren van de modellen, het onderhouden van de systemen etc. Hierbij is een open communicatie over de teams vaak niet noodzakelijk of wenselijk, zeker in het geval een team geen expertise over het onderwerp heeft. Wat dan wel van belang is, is een goede communicatie binnen het team. Bij Agile projecten waar vaak in een kleine teams alle projecttaken afgehandeld worden is een open communicatie zeer belangrijk. Hierbij past een agile projectmethodiek waarbij er volledige transparantie binnen het team nodig is om vroegtijdig problemen te onderkennen in het project (Batra, 2018; Saltz & Suthrland, 2019).

4.2.6. Theoretisch framework

In onderstaand framework is een samenvatting gegeven van de zojuist besproken raakvlakken van de projectkarakteristieken op de projectmethodiek thema's.

Data Science Project Framework			Iteratief			Maintenance	Conceptualization	Automate
			Agile-Plan balanced (Scrum)	Agile-Plan balanced (KanBan)	Agile-Heavy			
Project context	*Routinematige projecten*		x			x		x
	Eenmalige projecten				x			
Analytics context	*Hypothese testing*		x					
	Hypothese generating				x			
Data context	*Big Data*		x					
	Data-privacy		x				x	
Organisation Context	*Business ROI Culture*		x			x		
	Grote van de organisatie(onderdelen)	*Groot*	x					
		Klein			x			
Team Context	*Multidisciplinaire teams*		x	x				
	Open Communicatie				x			

Tabel 4 Thematisch raamwerk op basis van projectkarakteristieken

5. Demonstratie

De demonstratie heeft als doel om het theoretisch model, het artefact te toetsen op accuraatheid in de praktijk conform de Design Science methodiek. In dit onderzoek is dit gerealiseerd door het uitvoeren van een kwalitatief onderzoek d.m.v. een case studie bij data science teams die werkzaam zijn binnen twee organisatieonderdelen binnen een overheidsorganisatie. De data verzameling is uitgevoerd door te starten met een oriënterend documentonderzoek en vervolgens door het uitvoeren van interviews in twee iteraties. Als eerste zijn er interviews uitgevoerd met experts die binnen het domein analytics werkzaam zijn. Vervolgens zijn er focusgroep interviews uitgevoerd om dieper op het framework in te gaan, met als doel hierover te discussiëren en het framework te valideren. In de volgende paragrafen is beschreven met welk doel en op welke manier deze zijn uitgevoerd.

5.1. Document onderzoek

Voor het document onderzoek is een beeld gecreëerd van de domein architectuur analytics van de organisatie en de hierin voorgeschreven methodiek. De input voor dit proces was de 'Domein Architectuur Analytics'. Op basis van deze documentatie zijn er twee organisatieonderdelen geselecteerd met volwaardige data science teams. De volgende selectie criteria waren gesteld aan de teams op basis van het framework: Uitvoering van zowel innovatie als productgerichte data science projecten, diversiteit in omvang organisatieonderdeel, projectaanpak, beschikbaarheid stakeholders.

5.2. Expert interviews

Tijdens de eerste iteratie zijn vier expert interviews uitgevoerd met diverse experts uit de verschillende teams (tabel 5). Voor de expert interviews zijn d.m.v. een voorbereidend gesprek met een informant de belangrijke stakeholders geïdentificeerd en heeft geresulteerd in onderstaande lijst.

Expert Interviews
Lead Data Scientist, product team, organisatie onderdeel 1
Teamleider/projectleider, innovatie team, organisatie onderdeel 1
Lead Analytics Consultant, product team, organisatie onderdeel 2
Domein Architect Analytics, concern architectuur, organisatiebreed

Tabel 5 Participanten Expert Interviews

Het doel van deze iteratie is om een beeld te krijgen van de aanpak van de data science projecten binnen deze teams en organisatieonderdelen. De interviews waren semigestructureerd en op basis

van een vooral opgesteld interviewguide (bijlage 2) uitgevoerd. De interviews zijn via een online meeting uitgevoerd en duurden gemiddeld één uur per interview. De interviews zijn vervolgens getranscribeerd, gevalideerd (door de geïnterviewde), gecodeerd en geanalyseerd.
In onderstaande tabel is een overzicht gegeven van de organisatieonderdelen en de teams.

Organisatie onderdeel	Projectscope	Data Science methodiek	Projectmanagement methodiek	Data Science organisatie	Aantal teams
1	Routinematige projecten	CRISP-DM	Hybride (Scrum/KanBan)	- Klein, 15-20 medewerkers - Licht ROI gedreven cultuur	3
	Eenmalige projecten	CRISP-DM	Agile (Scrum/KanBan)	- Klein, 10-15 medewerkers - Geen directe management beïnvloeding	2
2	Routinematige en eenmalige projecten	CRISP-DM / Semma	Hybride (Scrum)	- Groot, 250 medewerkers - Sterk ROI gedreven cultuur	30

Tabel 6 Overzicht teams Case Studie

5.3. Focusgroep interviews

De tweede iteratie heeft plaatsgevonden door het uitvoeren van twee focusgroep sessies. Het voordeel hierbij is dat er niet gewerkt wordt met een interview vorm waarbij deelnemers één voor één antwoord geven op een vraag. De deelnemers worden aangemoedigd om met elkaar hierover te praten. Hierbij stellen deelnemers ook vragen aan elkaar en geven een reactie op elkaars ervaringen en meningen (Saunders, Lewis, & Thornhill, 2016). De interviewvragen zijn gebaseerd op de informatie uit de eerste iteraties en het framework. Elke focusgroep is uitgevoerd met drie participanten die werkzaam zijn in data science teams bij de organisatieonderdelen (tabel 7)

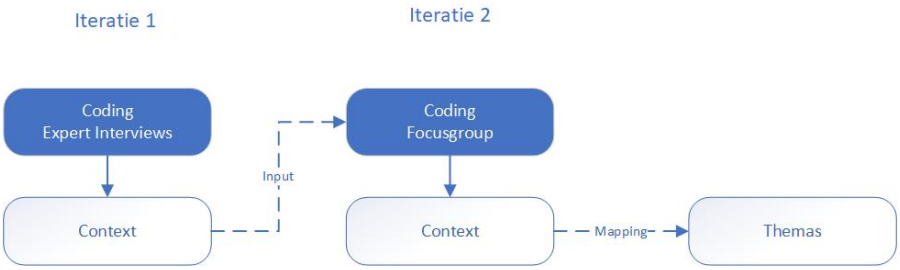
Focusgroep Interview – Organisatie onderdeel 1	Focusgroep Interview – Organisatie onderdeel 2
Lead Data Scientist/Business Analyst	Lead Data Scientist
Data Scientist	Business Analytics Consultant
Product Owner	Solution Architect Analytics

Tabel 7 Participanten Focusgroep Interview

Om het framework te toetsen is er gebruik gemaakt van een visueel model waarin de extra thema's t.o.v CRISP-DM zijn opgenomen. Tijdens de focus groep is er over de toevoegingen in het model gediscussieerd of deze extra stappen zinvol zijn en in welke context. Duidelijk is hier aangegeven dat het niet om het model gaat, maar over de toepassing van de concepten in het model. Het model, welke is gebruikt als discussieplaat tijdens de focusgroep, is toegevoegd in bijlage 4.

5.4. Transcriberen, valideren en coderen

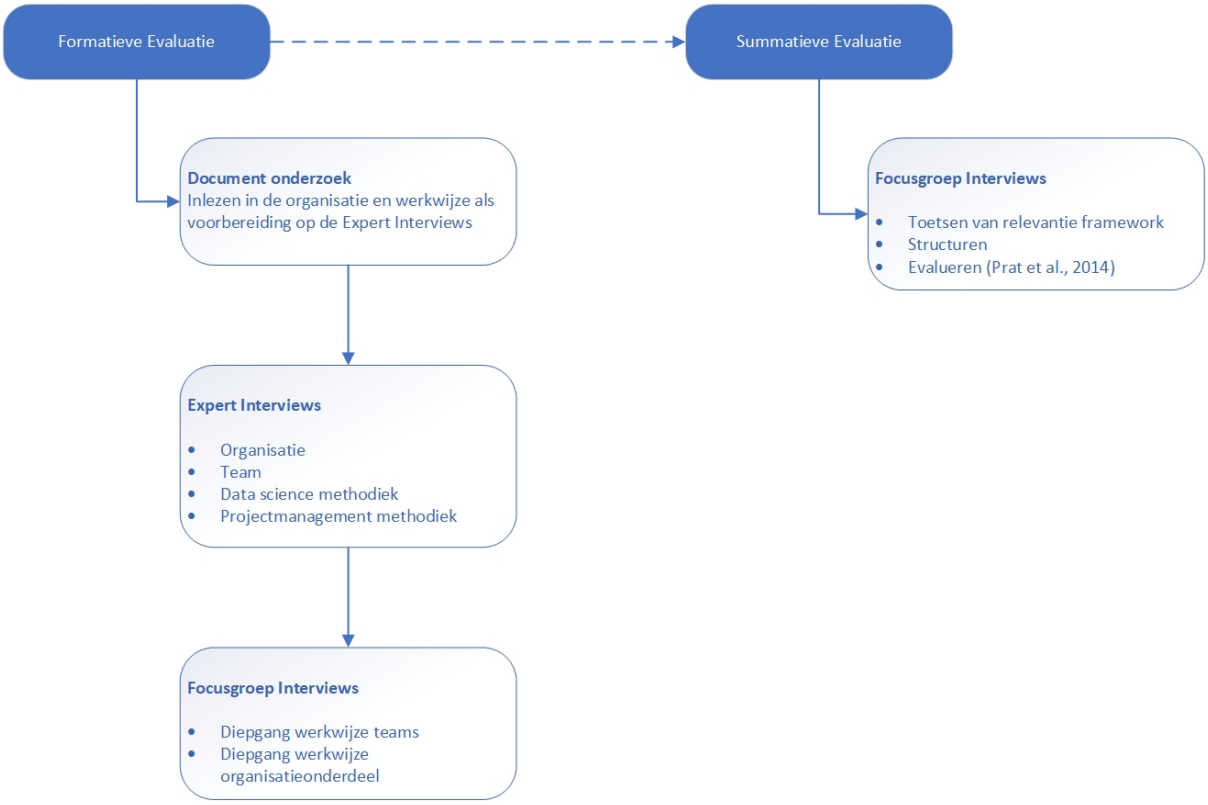
Alle interviews zijn getranscribeerd, geanonimiseerd en gevalideerd op correctheid door de geïnterviewden. Vervolgens zijn de transcripts gecodeerd met behulp van Atlas.Ti. Het coderen heeft plaatsgevonden in twee stappen: List coding en vervolgens selectief coderen. Voor de list coding is een code tabel opgesteld op basis van het framework met als doel om de informatie te structureren en te groeperen (bijlage 5). Tijdens het selectief coderen zijn de list codings toegekend aan de tekstfragmenten uit de transcripten. Deze tekstfragmenten zijn vervolgens gegroepeerd op basis van de codes en is de analyse gestart per code. Dit is zowel voor de expert interviews als voor de focusgroep sessie gehanteerd.



Figuur 9 Coding proces

6. Evaluatie

Dit hoofdstuk presenteert de resultaten van de formatieve en summatie evaluatie van het framework. Sectie 6.1 zal de resultaten presenteren van de formatieve evaluatie en sectie 6.3 van de summatieve evaluatie. In figuur 10 is de structuur van de evaluatie weergegeven.



Figuur 10 Structuur Evaluatie

6.1. Formatieve evaluatie

Voor de formatieve evaluatie is gestart met een documentonderzoek gevolgd door expert interviews en focusgroep sessies om de validiteit van de aspecten uit de design fase te valideren in de betreffende organisatie onderdelen.. De resultaten zijn in de volgende paragrafen beschreven.

Project context

De case organisatie heeft een domain architectuur Analytics die van toepassing is op alle data science teams binnen de organisatie. Deze architectuur schrijft voor om de BizDevOps cyclus (Chasioti, 2019) als leidraad voor routinematige projecten te hanteren. Deze cyclus borgt zowel de aspecten van de probleemformulering (BIZ), de data science methodiek (DEV) en het onderhoud en de automatisering (OPS).
Routinematige projecten worden binnen de case organisatie enkel agile plan-balanced uitgevoerd. Binnen een projectfase hanteert men voornamelijk de plannings en communicatie aspecten van Scrum. De probleemformulering, business understanding en data understanding fases zijn moeilijk in sprints in te plannen. Dit voornaamste oorzaak is dat de business vaak niet weet wat men wil en/of het verkrijgen van de benodigde data is een langdurig proces zeker bij privacy gevoelige data. Deze fasen worden vaak over meerdere sprints verspreid en men accepteert dat men niet altijd iets op kan leveren in een sprint. Alle teams geven aan dat de probleemformulering de belangrijkste fase uit het project is. Als het probleem niet duidelijk en smart geformuleerd is zal het product vrijwel nooit passen bij de behoefte van de business. De probleemformulering fase is zowel van toepassing voor routinematige als eenmalige projecten. Daarnaast wordt aangegeven dat onderhoud van het product essentieel is voor het succesvol uitrollen en gebruiken van het product. De belangrijkste fasen in deze maintenance cycle zijn model onderhoud, model performance en de business feedback tijdens de productie fase. Alhoewel de casus organisatie beperkt gebruik maken van automatisering geven zij aan dat dit aspect steeds belangrijker wordt. Vooral de automatisering van de ontwikkelstraat om sneller producten te kunnen opleveren en het automatisch testen om de kwaliteit tijdens het ontwikkelproces te borgen zijn belangrijke factoren die genoemd worden. Het BizDevOps model uit de architectuur bevestigt dit ook door de implementatie van een beheer cyclus.
Tussen de routinematige projecten wordt er tevens onderscheid gemaakt in de toegepaste projectmanagement methodiek als er zich een probleem voordoet welke vertragend gaat werken op de doorlooptijd van het project. Men gaat dan over van Scrum naar KanBan de focus ligt dan niet op wat er opgeleverd moet worden, maar op wat er opgeleverd kan worden. Opvallend is dat dit voornamelijk plaatsvindt binnen de teams bij de kleinere organisatieonderdelen die een grotere vrijheid hebben in het bepalen van hun werkwijze. Bij de doorontwikkeling van bestaande producten is de toepassing van Scrum en dan voornamelijk de sprint beter te plannen, doordat de data en het model er vaak al is, zijn de werkzaamheden vaak minder complex en beter te plannen in een sprint.
Voor de eenmalige vraagstukken werkt een planmatige aanpak niet. De doorlooptijden van deze type projecten variëren van een dag tot enkele weken en vragen om een hoog iteratieve agile aanpak. Vanwege het eenmalige karakter van deze type projecten is er geen behoefte aan onderhoud en automatisering van het ontwikkelproces. Er wordt voornamelijk een agile-heavy aanpak gehanteerd.

Project context	*Belangrijke thema's*
Routinematige projecten	Agile-plan balanced (Scrum), problem formulation, maintenance, automation
Eenmalige projecten	Agile-heavy, problem formulation

Tabel 8 Resultaten project context

Analytics context

De routinematige projecten werken hoofdzakelijk met hypothese testing analyses en hebben daarom vooral een agile-plan based aanpak en de data is beter te begrijpen voor de data scientist. Dit zijn veelal risicomodellen, waarbij gebruik gemaakt wordt van gestructureerde data met supervised modellen. Projecten met eenmalige vraagstukken werken vaker met hypothese genererende vraagstukken en gebruiken veelal unsupervised methodieken. Bij hypothese generating vraagstukken moet men veel vaker naar de business terug moet om de resultaten te kunnen begrijpen, de probleemformulering fase is daarom belangrijk om het probleem te begrijpen. Deze fasen worden wel in sprints opgepakt door alle teams, maar niet elke sprint wordt er iets opgeleverd voor de business. Het kan wel zijn dat bijvoorbeeld een bepaalde dataset het resultaat van een sprint is, dus een interne projectoplevering. Voor deze hypothese generating vraagstukken worden voornamelijk voor de eenmalige projecten een agile-heavy projectmethodiek toegepast.

Analytics context	*Belangrijke thema's*
Hypothese testing	Agile-plan balanced (Scrum)
Hypothese generating	Agile-heavy, problem formulation

Tabel 9 Resultaten analytics context

Data context

De teams werken met gestructureerde en ongestructureerde data. Deze data bestaat vaak uit data sets met een groot volume, maar is geen realtime data die snel verwerkt dient te worden. De karakteristieken van big data; volume, variety en velocity gaan hier niet op.
De organisatie werkt met gegevens van natuurlijke personen en rechtspersonen en zijn zeer privacy gevoelig. Het voorkomen van privacy schending en profilering is dus een belangrijk onderwerp bij de data science projecten. Hier zie je in de projectcyclus dat de data die gebruikt wordt eerst getoetst moet worden alvorens deze in het productieprocessen gebruikt mogen worden. Er moet hiervoor een soort conceptueel datamodel getoetst worden. Tijdens de modelering mag men wel meer data gebruiken, maar alvorens er een pilot mee uitgevoerd gaat worden moet het data model getoetst worden via een privacy impact analyse (PIA).
Tevens zie je in de casus organisatie dat privacy gevoelige data science vraagstukken, allemaal op een gestructureerde manier aangepakt worden in de reeds standaard gehanteerde Scrum projectaanpak (Agile-Plan balanced). Een getoetst conceptueel data model is vereist op het moment dat er met privacy gevoelige gegevens gewerkt gaat worden.

Data context	*Belangrijke thema's*
Data privacy	*Agile plan-balanced (Scrum), Conceptualization*
Big Data	

Tabel 10 Resultaten data context

Organisatie context

De organisatie is ROI gedreven voor de routinematige projecten en moeten planmatig werken met gebruik van agile principes. Elk project moet waarde opleveren voor de business, dit is zeker het geval voor het grote bedrijfsonderdeel met veel teams. Bedrijfsonderdelen met kleine teams hebben iets meer vrijheid en het innovatie team is niet ROI gedreven, maar de projecten zijn hier voornamelijk klein qua omvang en kort cyclisch. Bij het grootste bedrijfsonderdeel met 30 teams zien wij dat de organisatie een verplichte methodiek oplegt om een bepaalde kwaliteit te borgen in het proces. Hier wordt het Scaled Agile Framework (SaFe) gehanteerd. Dit is grotendeels op Scrum

gebaseerd en geschikt voor softwareontwikkeling voor een organisatie met veel teams. Wel zie je hier dat voornamelijk de bruikbare items uit de Scrum methodiek door de team gehanteerd worden. Bij de kleinere organisatieonderdelen zie je dat er meer vrijheid is in de keuze voor een projectmethodiek, voor routinematige projecten wordt veelal wel Scrum onderdelen gehanteerd, maar bij blokkerende problemen tijdens projecten gaat met over op KanBan en laat men de sprint prioriteiten en deliverables los. CRISP-DM zie je bij alle teams gehanteerd worden en wordt ook als de standaard gezien voor alle data science projecten. Binnen de eenmalige projecten wordt er meestal volledig agile gewerkt, dit is noodzakelijk door de korte doorlooptijden en urgentie van deze type vraagstukken.

Organisatie context	*Belangrijke thema's*
Business ROI Culture	Agile-plan balanced (Scrum), problem formulation, maintenance
Grote van de organisatie - klein	Agile-plan balanced (Scrum/ KanBan), Agile-heavy
Grote van de organisatie – groot	Agile-plan balanced (Scrum)

Tabel 11 Resultaten organisatie context

Team context

De teams geven allen aan dat zowel een open communicatie als multidisciplinaire teams essentieel zijn voor alle type projecten. Open communicatie en voornamelijk transparantie is belangrijk en voorkomt dat er producten gemaakt worden die de business niet wil of bij een team 'over de schutting' gegooid worden. Hiervoor hanteert men de communicatie momenten vanuit Scrum als de daily stand-up, sprint meetings, sprint evaluaties, demo's en sprint backlog. Door gebruik te maken van multidisciplinaire teams worden ook de relevante stakeholders betrokken bij de verschillende communicatie momenten, zodat iedereen op de hoogte is van de voorgang en de eventueel aanwezige problematiek. De teams hebben veelal een wisselende samenstelling gedurende de fases van het project. Op het moment dat een discipline nodig is word deze aan het team toegevoegd. De business is hier ook nauw bij betrokken, vaak in de vorm van een product owner die in het team geplaatst wordt. In deze teams zitten dus zowel afgevaardigden vanuit de business, maar ook data scientists, software ontwikkelaars en beheerders.

Team context	*Belangrijke thema's*
Multidisciplinaire teams	Agile-plan balanced (Scrum/ KanBan), Agile-heavy
Open Communicatie	Agile-plan balanced (Scrum/ KanBan), Agile-heavy

Tabel 12 Resultaten team context

6.2. Data Science Project Framework

Kijkende naar de formatieve evaluatie zijn er een aantal aspecten uit het initiële framework verwijderd of toegevoegd. De wijzigingen zijn in onderstaande tabel weergegeven en gevisualiseerd d.m.v. kleuren in het framework.

	Voor de verificatie	Na de verificatie
Problem formulation (thema) -> project context	Onderwerp was geen onderdeel van het framework. Werd wel in de literatuur benoemd, maar was geen key factor in eerste aanleg.	Tijdens de interviews blijkt dat hier de grootste knelpunten zaten voor alle teams. Daarom is dit thema wel significant voor het project succes.
Eenmalige projecten (project context)	Enkel Agile heavy als projectmethodiek	Door de innovatie teams werd er wel deels planmatig gewerkt om toch te kunnen sturen op een afronding van projecten, maar dan voornamelijk op basis van KanBan om inzicht in de hoeveelheid werk te houden.
Big Data (data context)	Big Data wordt agile plan balanced uitgevoerd, dus wel gestructureerde aanpak (CRISP-DM), maar meer iteraties tussen de fases.	De case organisatie werkt op dit moment niet met Big Data en kan dus niet getoetst worden in deze casus. Hier moet nader onderzoek voor verricht worden.
Business ROI culture (organisation context)	Agile plan-balanced aanpak met als primaire methodiek Scrum	In deze projecten is de waarde voor de business heel belangrijk dus een goed probleem formulatie en feedback is essentieel
Organisatie grootte (organisation context)	Kleine organisatieonderdelen werken voornamelijk agile heavy.	Innovatieve teams werken tevens planmatig en gebruikmakende van agile onderdelen van zowel Scrum als Kanban.
Multidisciplinaire teams (team context)	Belangrijk bij agile-plan balanced methodieken	Alle organisatieonderdelen werkten met multidisciplinaire team, dit wordt ale een belangrijk succes factor benoemd voor alle projecttypes.
Open communicatie (team context)	Belangrijk bij agile heavy	Alle organisatieonderdelen werkten hanteren een open communicatie, dit wordt als een belangrijke succes factor benoemd voor alle projecttypes.

Tabel 13 Aanpassingen framework na evaluatie

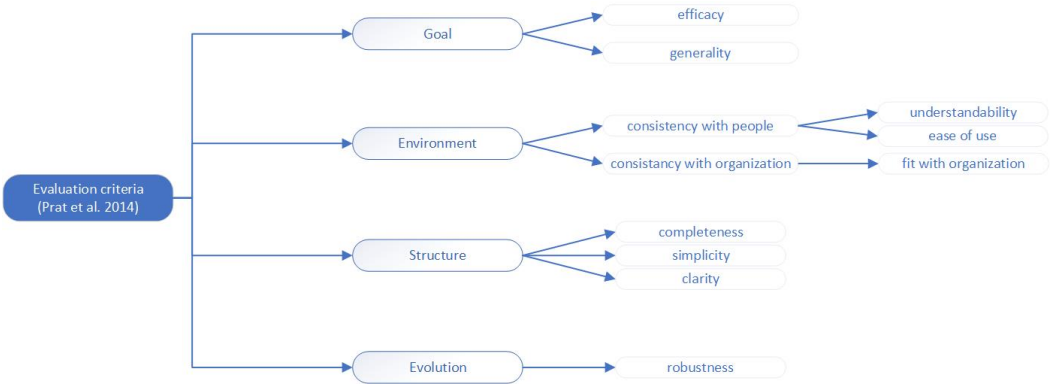
Dit heeft uiteindelijk geresulteerd in het onderstaande data science project framework.

Data Science Project Framework			Problem formulation	Iteratief			Maintenance	Conceptualization	Automate
				Agile-Plan balanced (Scrum)	Agile-Plan balanced (KanBan)	Agile-Heavy			
Project context	*Routinematige projecten*		x	x			x		x
	Eenmalige projecten		x		x	x			
Analytics context	*Hypothese testing*			x					
	Hypothese generating		x			x			
Data context	*Big Data*								
	Data-privacy			x				x	
Organisation Context	*Business ROI Culture*		x	x			x		
	Grote van de organisatie(onderdelen)	*Groot*		x					
		Klein		x	x	x			
Team Context	*Multidisciplinaire teams*			x	x	x			
	Open Communicatie			x	x	x			

Tabel 14 Data Science Project framework na evaluatie fase (groen = toevoeging, oranje = verwijderd)

6.3. Summatieve evaluatie

De summatieve evaluatie op het framework is uitgevoerd door twee focusgroep interviews bij de twee organisatieonderdelen uit te voeren. De focusgroep sessies hadden als doelstelling om meer diepgang op het framework te krijgen en de contexten en de relatie met de thema's uit het framework te toetsen op relevantie en bruikbaarheid in de praktijk.
Het data science project framework is vervolgens geëvalueerd op basis van de evaluatie criteria van Prat et al. (2014). Tijdens deze studie was er geen ruimte om het model daadwerkelijk toe te passen in de praktijk om op deze manier een beeld te krijgen van de werking en performance van het framework, daarom zijn de evaluatie criteria 'activity' en 'utility' niet geëvalueerd tijdens deze studie.



Figuur 11 Evaluatie criteria

Goal
Tijdens de focusgroep sessies is de relevantie en toepasbaarheid in de projecten bevestigd door de participanten. Het framework beschrijft precies de thema's waar de diverse teams mee worstelen tijdens de projecten en geven richting aan de manier hoe hier mee om te gaan. De matching van de diverse projectkarakteristieken aan de thema's worden in de focusgroep sessies bevestigd en essentieel bevonden. Hierdoor is het framework toepasbaar op de verschillende type data science projecten. Er wordt in de sessie bevestigd dat data science projecten op een andere manier aangepakt moeten worden dan de traditionele software enginering projecten, de componenten uit het framework en de relaties tussen de diverse factoren en thema's dragen hieraan bij.

Environment
Alhoewel het framework niet letterlijk getoond is aan de participanten van de focusgroep sessies, zijn tijdens deze sessie wel de begrippen, de relaties en de toepasbaarheid in de case organisatie getoetst door gebruik te maken van een fictief model (bijlage 4) die is opgebouwd uit een samenvoeging van de thema's uit het framework en de processtappen van CRISP-DM. De diverse thema's zijn besproken, nader uitgelegd indien noodzakelijk en vervolgens getoetst, geverifieerd en bevestigd op toepasbaarheid in de praktijk door de participanten.

Structure
Tijdens de focusgroep sessie zijn de begrippen uit het framework besproken en zijn deze in detail bediscussieerd op basis van een fictief model met de thematische aspecten van het framework. Voornamelijk het begrip conceptualisatie was niet direct duidelijk voor alle participanten, maar na een verduidelijking is er consensus over de betekenis en het doel van dit thema. Daarnaast is besproken of er essentiële onderdelen missen die vanuit hun expertise invloed hebben op de project aanpak, dit was niet het geval.

Hieruit kan de conclusie getrokken worden dat de samenhang tussen de projectcontext en de thema's valide is.

Evolution
De doelstelling van het framework is om de projectaanpak aan te passen op de diverse type projecten en moet dus robuust genoeg zijn om hiermee om te kunnen gaan. Door de informatie uit zowel de expert interviews en focusgroep sessies kan geconcludeerd worden dat de diverse projectkarakteristieken en de hierbij belangrijke thema's volledig en compleet waren, dit bevestigt de robuustheid van het framework en dus de primaire aspecten die van belang zijn hierin benoemd worden.

7. Conclusies en aanbevelingen

In dit hoofdstuk is de conclusie beschreven op basis van de resultaten van het design science onderzoek dat is uitgevoerd in deze thesis. Na de conclusies wordt er verder ingegaan op de implicaties voor de wetenschap en practitioners en zal aanbevelingen doen voor verder onderzoek.

7.1. Conclusies

Het theoretisch framework (figuur 4) wat is ontstaan uit het literatuuronderzoek beantwoordt de deelvragen van dit onderzoek. Het design science onderzoek had als doel om een framework te realiseren waarmee organisaties een projectaanpak kunnen creëren die aansluit bij het data science vraagstuk. Door het ontwikkelde framework is er inzicht in de factoren die de aanpak van een data science project beïnvloeden en vervolgens welke thema's relevant zijn voor de projectaanpak. Hiermee kan voor data science projecten bepaald worden welke cruciale stappen men in de projectcyclus moet toevoegen om het project succesvol te doorlopen en een product op te leveren waar de business om gevraagd heeft. Daarmee beantwoord het framework in deze thesis de primaire onderzoeksvraag:

'Hoe beïnvloeden projectkarakteristieken de keuze voor een succesvolle projectmethodiek voor data science projecten?'

Als we naar het framework (tabel 14) kijken kunnen we de volgende conclusies trekken:
De belangrijkste onderscheidende factor bij projecten is de project context deze bepaalt in hoofdlijnen de aanpak van een project. Bij de routinematige projecten is (1) een planmatige aanpak met agile principes de leidraad. Scrum principes zijn hier de standaard, maar het principe dat elke sprint een product oplevert is los gelaten tijdens de probleemformulering, business en data understanding fase. Tevens is (2) maintenance essentieel voor de lifecycle van het model en moeten in een project opgenomen worden en (3) draagt automatisering bij aan de kwaliteit en snelheid van opleveren door testen en deployment te automatiseren. Daarentegen worden eenmalige projecten, vaak hypothese generating van aard in een hoge mate agile uitgevoerd, waarbij er vaak gebruik gemaakt wordt van KanBan en Scrum om zicht op de hoeveelheid werk te houden. ROI gedreven projecten in grote organisaties zijn vaak routinematig en hypothese testing van aard, hierbij wordt vanwege het afbreukrisico een agile-plan balanced aanpak gehanteerd om control op het project te houden. Bij kleine organisatie (onderdelen) is er meer vrijheid voor de teams in de projectmanagement methodiek. Bij elk project moet een open communicatie de standaard zijn. Drijfveer hiervoor is dat het project transparant moet zijn voor alle stakeholders, zodat er direct bijgestuurd kan worden als er afwijkingen geconstateerd worden. Multidisciplinaire teams zorgen

ervoor dat iedere stakeholder op het juiste moment betrokken blijft en input kan leveren voor het project.

7.2. Implicaties voor de wetenschap

De Design Science Research van dit onderzoek voegt drie contributies aan de wetenschap van information science toe.
Als eerste, heeft het literatuuronderzoek inzicht gegeven in de projectkarakteristieken die invloed hebben op de aanpak van data science projecten door het uitvoeren van een literatuuronderzoek, waarmee de diverse onderkende karakteristieken uit de diverse papers is samengebracht in dit framework. Ten tweede heeft het literatuuronderzoek in beeld gebracht hoe de hedendaagse data science methodieken zich onderscheiden van CRISP-DM en dit vertaalt in thema's die van belang zijn voor het uitvoeren van succesvolle data science projecten. Deze thema's zijn vastgesteld op basis van overeenkomstige verschillen uit CRISP-DM geëvolueerde methodieken en modellen.
Als derde heeft dit geresulteerd in een artefact, het framework, welke op basis van de projectkarakteristieken in een project aanbevelingen doet voor de projectaanpak. Dit framework kan daarom als basis dienen voor verder wetenschappelijk onderzoek.
Deze studie heeft geresulteerd in een data science project framework framework dat getoetst is op bruikbaarheid tijdens een case studie onderzoek. Een kwalitatieve studie is noodzakelijk om het framework te toetsen op werking en performance door deze in data science projecten toe te passen.

7.3. Aanbevelingen voor de praktijk

Het Data Science project framework is uitvoerbaar en bruikbaar in de praktijk. Dit is bevestigt tijdens de focusgroep sessies. Het framework is wellicht niet getest op effectiviteit en performance tijdens dit onderzoek, maar de inhoud van het framework is relevant, essentieel en toepasbaar in de praktijk. Om effectiviteit van het framework te toetsen is de aanbeveling om te starten bij de organisatieonderdelen met de kleinere data science teams die meer vrijheid hebben in de keuze van hun werkwijze. Het is van belang om de diverse projecttypen te groeperen om zo een beeld te krijgen van het type projecten, de analyse vraagstukken en de type gegevens. Plot deze typen, rekening houdende met de team en organisatie context, op het framework en selecteer op basis van deze resultaten de relevante thema's uit het framework en implementeer deze vervolgens in de projectmanagement methodiek. Laat vooral in de beginfase van het project tijdens de *probleem formulering*, de *business en de data understanding* de strakke sprintplanningen los, maak hier een globale schatting van de hoeveelheid werk op basis van ervaringen uit andere projecten en plan deze in. Dit kan dus betekenen dat deze stappen variëren van een enkele dag tot enkele maanden, accepteer dat in deze fase van het project. Hanteer wel de sprintmeetings en daily stand-ups om zicht te houden op de voortgang en het ontstaan van eventuele problemen.

7.4. Limitatie en verder onderzoek

De volgende limitaties zijn van toepassing op dit onderzoek:

- De expert interviews hebben plaatsgevonden met vier experts uit het data science werkveld, van zowel product georiënteerd werkende teams, een innovatief georiënteerd team en een domein architect om een beeld te krijgen van de huidige projectaanpak, organisatieonderdelen, team samenstelling en de aanwezige problematiek. Tijdens de focusgroep interviews konden niet al deze disciplines aanwezig zijn en hebben deze sessies zich beperkt tot voornamelijk data scientists en analisten. Het missen van deze stakeholders kan impact hebben gehad op de validiteit en kwaliteit van het onderzoek. Als deze wel zouden kunnen aansluiten bij de focusgroep sessies zouden dit impact gehad kunnen hebben op de evaluatie van het framework.

Een gevarieerde samenstelling van de focusgroepen zouden wellicht een geresulteerd hebben in een mindere mate van bias in het onderzoek.
- Het Data Science Project Framework is niet getoetst op werking in de praktijk. Dus op een daadwerkelijke implementatie in de data science projecten. Een dergelijke toetsing van het framework in de praktijk zou de validiteit en kwaliteit van het framework bevestigen.
- Het onderzoek is enkel uitgevoerd bij een overheidsorganisatie die ROI gedreven is en niet werkt met big data vraagstukken. Hierdoor kunnen niet alle factoren uit het initiële theoretisch framework getoetst worden (Big Data) en geeft het onderzoek een beeld van enkel één type organisatie.

Kijkende naar de limitatie van dit onderzoek kunnen we de volgende aanbevelingen doen voor verder onderzoek:
- Nader onderzoek is nodig voor de implementatie van het framework in de praktijk om deze te toetsen op effectiviteit en performance.
- Nader onderzoek is noodzakelijk voor de toepassing van het framework bij andere type organisaties (bijv. commerciële organisaties) en organisaties die big data projecten uitvoeren.

Deze aanbevelingen zullen bijdragen aan een robuust framework die generiek toepasbaar is voor projecten in het data science werkveld.

Referenties

Ahangama, S., & Poo, D. C. C. (2015a). *Designing a Process Model for Health Analytic Projects.* Paper presented at the PACIS.

Ahangama, S., & Poo, D. C. C. (2015b). *What Methodological Attributes Are Essential for Novice Users to Analytics?–An Empirical Study.* Paper presented at the International Conference on Human Interface and the Management of Information.

Anderson, D. J. (2010). *Kanban: successful evolutionary change for your technology business*: Blue Hole Press.

Asadi Someh, I., Breidbach, C. F., Davern, M. J., & Shanks, G. (2016). ETHICAL IMPLICATIONS OF BIG DATA ANALYTICS. In *Research-in-Progress Papers*.

Bach, M. P., Zoroja, J., & Celjo, A. (2017). An extension of the technology acceptance model for business intelligence systems: project management maturity perspective. *Ijispm-International Journal of Information Systems and Project Management, 5*(2), 5-21. doi:10.12821/ijispm050201

Baijens, J., & Helms, R. W. (2019). Developments in knowledge discovery processes and methodologies: anything new?

Batra, D. (2017). Adapting Agile Practices for Data Warehousing, Business Intelligence, and Analytics. *Journal of Database Management, 28*(4), 1-23. doi:10.4018/JDM.2017100101

Batra, D. (2018). Agile values or plan-driven aspects: Which factor contributes more toward the success of data warehousing, business intelligence, and analytics project development? *Journal of Systems and Software, 146*(1), 249-262. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eoah&AN=46651558&site=ehost-live

Beck, K., Beedle, M., Bennekum, A. v., Cockburn, A., Cunningham, W., Fowler, M., . . . Thomas, d. (2001). Manifest voor Agile Software Ontwikkeling. Retrieved from https://agilemanifesto.org/iso/nl/manifesto.html

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 16*.

Chasioti, K. (2019). *BizDevOps: A process model for the Alignment of DevOps with Business Goals.*

Das, M., Cui, R., Campbell, D. R., Agrawal, G., & Ramnath, R. (2015). *Towards methods for systematic research on big data.* Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).

Algemene Verordening Gegevensbescherming (EU) 2016/679, (2016).

Franková, P., Drahošová, M., & Balco, P. (2016). Agile Project Management Approach and its Use in Big Data Management. *Procedia Computer Science, 83*(1), 576-583. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eoah&AN=38939501&site=ehost-live

Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects.

Gartner. (2018). Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

Grady, Payne, J., & Parker, H. (2017, 11-14 Dec. 2017). *Agile big data analytics: AnalyticsOps for data science.* Paper presented at the 2017 IEEE International Conference on Big Data (Big Data).

Grady, N. (2016). *Knowledge Discovery in Data Science.* Paper presented at the Proc. IEEE International Conference on Big Data.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems, 19*(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

IBM. (2016). Extracting business value from the 4 V's of Big Data. Retrieved from https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal, 80*(4), 463-475.

Li, Y., Thomas, M. A., & Osei-Bryson, K.-M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems, 91*(1), 1-12. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eoah&AN=39655600&site=ehost-live

Marbán, Ó., Mariscal, G., Menasalvas, E., & Segovia, J. (2007). *An engineering approach to data mining projects.* Paper presented at the International Conference on Intelligent Data Engineering and Automated Learning.

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25*(2), 137-166.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems, 24*(3), 45-77.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). *Artifact Evaluation in Information Systems Design-Science Research-a Holistic View.* Paper presented at the PACIS.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*: " O'Reilly Media, Inc.".

Richarz, A. N., Vessela, V., Kochev, N., Myatt, G., Jeliazkova, N., Neagu, D., . . . Vracko, M. (2019). *Big Data in Predictive Toxicology*: Royal Society of Chemistry.

Saltz, J. (2015). *The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness.* Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).

Saltz, J. (2017). ACCEPTANCE FACTORS FOR USING A BIG DATA CAPABILITY AND MATURITY MODEL. In *Research-in-Progress Papers* (pp. 2602-2612 %U https://aisel.aisnet.org/ecis2017_rip/2613).

Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018). Exploring Project Management Methodologies Used Within Data Science Teams.

Saltz, J., & Shamshurin, I. (2015). *Exploring the process of doing data science via an ethnographic study of a media advertising company.* Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).

Saltz, J., & Shamshurin, I. (2016, 5-8 Dec. 2016). *Big data team process methodologies: A literature review and the identification of key factors for a project's success.* Paper presented at the 2016 IEEE International Conference on Big Data (Big Data).

Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology, 68*(12), 2720-2728.

Saltz, J., & Suthrland, A. (2019). *SKI: An Agile Framework for Data Science.* Paper presented at the 2019 IEEE International Conference on Big Data (Big Data).

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research Methods For Business Students* (Seventh edition ed.): Pearson Education Limited.

Schwaber, K., & Sutherland, J. (2017). The Scrum GuideTM. The definitive guide to scrum: The rules of the game. November 2017. In.

Siddiqa, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications, 71*, 151-166. doi:https://doi.org/10.1016/j.jnca.2016.04.008

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European journal of information systems, 25*(1), 77-89.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.

Bijlage 1 – Literatuur review procedure

Om een duidelijke beeld te krijgen van het probleem zijn er een drietal deelvragen geformuleerd in Hoofdstuk 1 van dit onderzoek, namelijk:

D1 'Hoe onderscheiden de hedendaagse projectmethodieken binnen het data science werkveld zich van elkaar?' (literatuuronderzoek)
D2 'Welke projectkarakteristieken hebben invloed op de aanpak van data science projecten?' (literatuuronderzoek)
D3 'Welke invloed heeft een projectmanagement methodiek op het succesvol afronden van een data science project?' (literatuur- en kwalitatief onderzoek)

Het literatuuronderzoek heeft als doel om deze deelvragen te beantwoorden en om een theoretisch overzicht te generen van de reeds bestaande kennis met als doel om een artifact te ontwikkelen voor dit onderzoek. De aanpak van dit literatuuronderzoek is hieronder in detail uitgewerkt.

Planning

- *Protocol*

Het van te voren vastleggen van een protocol is volgens Okoli and Schabram (2010) een cruciale stap voor het verdere literatuuronderzoek. Voor het formuleren van het protocol is de Systematic Literature Research (SLR) methodiek gehanteerd van Okoli and Schabram (2010). Met behulp van deze methodiek is er een protocol geformuleerd dat gestructureerd inzicht geeft in het doorlopen proces en geeft de mogelijkheid tot reproduceerbaarheid om de externe validiteit van het onderzoek te borgen. In het onderstaande diagram is het gevolgde protocol voor de uitvoer van het literatuuronderzoek weergegeven.

Figuur 12 Onderzoeksprotocol literatuuronderzoek op basis van SLR

Selectie

- *Start-set*

De Open Universiteit heeft voor dit onderzoek een eerste set met wetenschappelijke artikelen beschikbaar gesteld om kennis te nemen van het te onderzoeken fenomeen en de begrippen die in dit werkveld voorkomen. Deze artikelen vormen de start-set voor dit onderzoek.

Artikelnr	Artikel
A1	Ahangama, S., and Poo, D. C. C. 2015. "What Methodological Attributes Are Essential for Novice Users to Analytics? - an Empirical Study," in International Conference on Human Interface and the Management of Information (Vol. 9173), pp. 77–88.
A2	Mariscal, G., Marbán, Ó., and Fernández, C. 2010. "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies," *Knowledge Engineering Review* (25:2), pp. 137–166.
A3	Li, Y., Thomas, M. A., and Osei-Bryson, K.-M. 2016. "A Snail Shell Process Model for Knowledge Discovery via Data Analytics," Decision Support Systems (91), Elsevier B.V., pp. 1–12. (https://doi.org/10.1016/j.dss.2016.07.003).

A4	Saltz, J., Shamshurin, I., and Connors, C. 2017. "Predicting Data Science Sociotechnical Execution Challenges by Categorizing Data Science Projects," *Journal of the Association for Information Science and Technology* (68:12), pp. 2720–2728. (https://doi.org/10.1002/asi.23873).
A5	Saltz, J. S., Wild, D., Hotz, N., and Stirling, K. 2018. "Exploring Project Management Methodologies Used Within Data Science Teams," in *Twenty-Fourth Americas Conference on Information Systems, New Orleans, 2018*, pp. 1–5.
A6	Baijens, J., & Helms, R. W. (2019). Developments in knowledge discovery processes and methodologies: anything new?

Tabel 15 Initieel beschikbaar gestelde artikelen door de OU

- *Conceptuele begrippen*

Op basis van de start-set is er een lijst met concepten geformuleerd die als basis dient voor de zoekopdracht naar gerelateerde artikelen in de diverse wetenschappelijk databases. Het concept is in de lijst opgenomen als deze vaker voorkomt in de titel, het abstract en de inleiding van het artikel uit de start-set. In onderstaande tabel zijn deze concepten (key terms) weergegeven. Als hulpmiddel voor het gestructureerd vastleggen van de begrippen is gebruik gemaakt van mindmapping.

	Conceptuele begrippen (AND)	
gerelateerde begrippen (OR)	Data science	Project methodology
	Data analytics Knowledge discovery Business intelligence Data mining Big Data	Project management Project framework Agile Process methodology Process model

Tabel 16 Conceptuele begrippen en gerelateerde begrippen t.b.v. literatuuronderzoek

- *Selectie van wetenschappelijke databases en query*

Er zijn voor dit onderzoek vier databases geraadpleegd, namelijk ACM, AIS, EBSCOhost en Web of Science. Deze vier databases zijn geselecteerd, omdat deze een grote database aan artikelen over de onderzoeksgebieden Informatica en Informatiesystemen bevatten. Er is één generieke query geformuleerd met de begrippen uit tabel 2. Deze query is uniform toegepast is in de vier geselecteerde databases. De query en de gevonden resultaten per database zijn opgenomen in tabel 3.

Generieke Query	
(("data science" OR "data analy*" OR "knowledge discover*" OR "business intelligence" OR "data mining" OR "big data") AND ("project method*" OR "project management" OR "project framework*" OR "process method*" OR "process method*" OR "process model*" OR "agile"))	
Database	Resultaten
ACM	111
AIS	20
EBSCOhost	97
Web of science	42
	Totaal
Totaal aantal artikelen	270
Na ontdubbeling	197
Na selectie op relevantie	27
Na Quality appraisal	13
Toegevoegd voor snowballing	8
Totaal geschikt voor de thesis	21

Tabel 17 Database query's t.b.v. literatuuronderzoek

Van alle databases zijn van alle artikelen de auteur(s), titel, jaartal en abstract geëxporteerd in RIS formaat voor verdere analyse op relevantie.

- *Eerste selectie op basis van relevantie (practical screening)*

Op de eerste set met artikelen heeft als eerste ontdubbeling plaatsgevonden. Hierna bleven er 197 artikelen over. Van deze 197 artikelen is vervolgens een selectie gemaakt op basis van relevantie met de onderzoeksvraag. De relevantie van de artikelen is gescreend op het feit dat deze een relatie hadden met data science in relatie tot projectaanpak, project problematiek en methodieken. Technische artikelen die ingaan op data science modellen vielen bijv. buiten de scope van de selectie. De selectie is gemaakt op basis van de titel van het artikel, aangevuld met het abstract daar er twijfel aanwezig was voor de relevantie. Na deze selectie blijven er 27 artikelen over.

Extractie

- *Quality Appraisal*

Alle 27 overgebleven artikelen doorgaan een verdere selectie op relevantie op de onderzoeksvragen op basis van het lezen van de inleidingen en conclusies van de artikelen. Daarnaast is gekeken of de wetenschappelijke artikelen peer-reviewed zijn. Deze analyse resulteert in een uiteindelijke selectie van 13 artikelen. Door snowballing komen er nog 8 artikelen bij, wat het totaal op 21 artikelen brengt.

- *Data Extractie*

De uiteindelijke 21 artikelen zijn in het geheel doorgenomen en de informatie is verwerkt en gestructureerd in een mindmap en in samenvattingen van de artikelen. De relevante informatie voor de onderzoeksvragen is opgenomen in het theoretisch kader van deze thesis en gestructureerd op thema.

Uitvoering

- *Schrijven theoretisch kader*

Het theoretisch kader is uitgewerkt in H2.2 Resultaten en conclusies.

Bijlage 2 Interviewguide Iteratie 1 - Expert interviews

Opening

- Dank voor de acceptatie en deelname aan dit interview.
- Introductie van het doel van het onderzoek en uitleg van het doel van dit interview.
- Het interview wordt getranscribeerd en geanonimiseerd.
- Het interview wordt opgenomen, enkel voor de transcriptie, na verwerking en goedkeuring van de transcriptie wordt de opname verwijderd.

Thema's van het interview

Het doel is om een diepgaand beeld te krijgen van de werkwijze van de organisatie betreffende de aanpak van data science projecten. Met als doel om tijdens de discussie met iedereen een diepgaand inzicht gekregen wordt in de problematiek van de huidige werkwijze en ook of dit beeld door alle participanten gedeeld wordt.

Organisatie context

De organisatie kan een rol spelen in de wijze waarop data science projecten uitgevoerd worden. Is de organisatie business case georganiseerd? Is er een vooraf gedefinieerd proces? Heeft de omvang van de organisatie impact? Etc.

1. *Hoe ziet jullie organisatie eruit? (organisatie context)*
2. *Wat is de verwachting van de organisatie van jullie projecten? (organisatie context)*
3. *In welke mate beïnvloed de organisatie of je een project kunt/mag oppakken. (organisatie context)*
4. *Hoe bepalen jullie of een opdracht haalbaar is? (probleem formulering)*
   Naar welke aspecten kijken jullie dan?
5. *Hoe wordt de organisatie betrokken bij de resultaten van de projecten? (probleem formulering)*

Team context

De team context heeft impact op de aanpak van de data science projecten. Hoe zijn de teams opgebouwd? Is dit een vaste samenstelling?

1. *Hoeveel data science teams hebben jullie? (team context)*
2. *Hoe zijn deze teams samengesteld? (team context – multidisciplinaire teams)*
3. *Waarom hebben jullie gekozen voor deze samenstelling? (team context – multidisciplinaire teams)*
4. *Hoe beïnvloed het project de samenstelling van de teams? (team context – multidisciplinaire teams)*
5. *Hoe belangrijk is communicatie in het team en hoe borgen jullie dit? (team context – open communicatie)*

Project context

Met deze vragen wil ik een beeld krijgen van de type projecten, de aanpak en factoren welke deze aanpak beïnvloeden.

1. *Welke soorten verschillende projecten voeren jullie uit (project context)?*
   In welke categorieën kun je deze onderverdelen? Wat moet het opleveren uiteindelijk.
2. *Waar onderscheiden de verschillende projecten zich voornamelijk in? (project context)*
3. *Hoe pakken jullie de projecten nu aan? (procesmodel)*
   Kun je beschrijven hoe jullie de verschillende type data science projecten uitvoeren, vanaf de klantvraag tot oplevering van het resultaat en verder? (procesmodel)

4. *Welke factoren bepalen voornamelijk de gekozen aanpak van de projecten? (project context)*
5. *Hoe beïnvloeden de verschillende projectentypes de gehanteerde projectmethodiek ? (procesmodel- diversiteit in aanpak)*
6. *Worden projecten waterval of agile aangevlogen of juist in een combinatie van beide projectmanagement methodieken? (procesmodel - iteratief)*
7. *Wanneer is een project afgerond? (procesmodel - maintenance)*
8. *Hoe lang duren de projecten? Hoe verklaar je de verschillen in de duur?*
9. *Hoe stellen jullie de projectresultaten/producten beschikbaar aan jullie opdrachtgevers? (procesmodel – automation)*
10. *Hoe onderhouden jullie de producten? (procesmodel – maintenance)*
11. *Hoe borgen jullie dit onderhoud structureel? Of is dit niet altijd nodig? (procesmodel – maintenance)*

Analytics context

De analytische context omvat de analyses die uitgevoerd moeten worden door de teams. Specifiek het type analyse.

1. *Welke type analyses voeren jullie uit? (analytics context -hypothese testing/generating)*
2. *Hoe heeft het type analyses invloed op de te hanteren projectmethodiek? (analytics context - procesmodel)*

Data context

De data context maakt onderscheidt tussen big data (kunt er veel informatie in vinden, maar resultaat is niet altijd bekend) en small data (weet van te voren welke resultaten er zijn) betreffende de project aanpak. Ook de data privacy heeft impact op de manier van werken.

1. *Wat voor type data kom je tegen in jullie projecten? (data context – big or small data) (steekwoorden: gestructureerd, ongestructureerd, big data, small data)*
2. *Hoe beïnvloed deze data de aanpak van je project? (procesmodel)*
3. *Hoe speelt (data) privacy een rol in jullie projecten? (data context - data privacy) Hoe pak je dit soort aspecten aan? (procesmodel)*
4. *Hoe borg je de data privacy gedurende je project lifecycle en daarna? (data context – procesmodel – maintenance)*

Bijlage 3 Interviewguide Iteratie 2 – Focusgroep

Deze guide is een handvat voor het doorlopen van het virtuele discussiemodel uit bijlage 4 en heeft als doelstelling de rode lijn te blijven hanteren.

Opening

- Dank voor de acceptatie en deelname aan dit interview.
- Introductie van het doel van het onderzoek en uitleg van het doel van dit interview.
- Het interview wordt getranscribeerd en geanonimiseerd.
- Het interview wordt opgenomen, enkel voor de transcriptie, na verwerking en goedkeuring van de transcriptie wordt de opname verwijderd na afronding van de thesis.

Thema's van het interview

Het doel is om een diepgaand beeld te krijgen van de werkwijze van de organisatie betreffende de aanpak van data science projecten. Met als doel om tijdens de discussie met iedereen een diepgaand inzicht gekregen wordt in de problematiek van de huidige werkwijze en ook of dit beeld door alle participanten gedeeld wordt. Als discussie wordt een procesmodel gehanteerd. Dit model heeft als doel om het framework te toetsen op bruikbaarheid en is heeft niet als doel ….

Problem formulation

Doel om het business probleem helder te krijgen dat het project moet oplossen en deze te transformeren naar een bruikbare analytische probleemstelling (hypothese). Meest belangrijke stap uit het besluitvormingsproces.

Doelstellingen

- Determine business objectives and success measures.
  Duidelijkheid van het doel van de business en hoe je dit kunt meten
- Deploy problem formulation strategies.
  Bepaal de grenzen, splits complexe problemen op in kleinere problemen.
- Define business problem.
  Er is een probleem definitie geformuleerd die een antwoord geeft op de wat, waarom en hoe vragen.
- Determine KDDA problem, goals, and success measures.
  Het type data science vraagstuk wordt bepaald op basis van het bedrijfsprobleem en de doelstellingen. Definieer hoe je de analytische doel kunt gaan meten (evaluatie)

1. Hoe vind dit proces nu plaats met de business om het probleem helder te krijgen? (Product Owner)
2. Welke methodiek hanteren jullie om de problemen vanuit de business helder en concreet te krijgen? (SMART, opsplitsing in kleine problemen)
3. Hoe formuleren jullie het business probleem? (methodiek)
4. Is dit ook een proces/model dat altijd bij elk data science vraagstuk gehanteerd wordt? Of is er onderscheid tussen de vraagstukken te maken?

DATA CYCLE

Business Understanding (CRISP-DM)

Helder krijgen van de requirements, formuleer de Business Case (costs and benefits), analytics capability, determine the PM method (waterfall, agile, mixed).

- Organizational analytics maturity (Analytics environment in the organization)
- Data maturity (Data suitable for analytics?)
- Decision style maturity (business users decision style mature enough?)

Data Understanding (CRISP-DM)

- De data leren begrijpen (structuur, grote en formaat)
- Inzetten van visualisatie om bijv. big data te begrijpen.
- Data beschrijving; karakteristieken, bronsystemen, update frequentie, kwaliteit etc.

1. Wat voor type data komen jullie tegen? Gestructureerd, ongestructureerd
2. Hoe beïnvloed het type data (big data vs small data) jullie projectaanpak?
3. Heeft het type dat effect op de doorlooptijd van een project?

Conceptualization (USAM)

Conceptualizatie heeft als doel om een conceptueel datamodel te maken op basis van de onderzoeksvragen. Hieronder valt ook de theorie die gebruikt is om het model te creeren, een beschrijving van de variabelen in het model (afhankelijke en onafhankelijke variabelen).
Het is beter om zinvolle variabelen te hanteren die je kunt verantwoorden op basis van ervaring en literatuur.

Deployment & Automation

Naast de deployment van het model naar productie is het van belang om

1. Hoe vind jullie deployment van de modellen plaats?
2. Hebben jullie items in jullie proces geautomatiseerd? Waar in het proces vindt dit plaats
3. Heeft het type project hier nog invloed op?
4. Waarom heb je dit wel of niet gedaan?
5. Wat is de noodzaak om te automatiseren?

Maintenance

- Moeilijke aansluiting met de business.
- Beginfase van data sciencefiction project moeilijk in strakke sprints te plannen. (Business Understanding/ Data Understanding)
- Maintenance/Life Cycle(data/modellen) is noodzakelijk, maar nog niet concreet ingericht.

1. Heeft elk project een maintenance cycle (life-cycle)? (Of zou dat moeten hebben)
2. Waarom is dit wel/niet zo?
3. Welke stappen zien jullie in de life-cycle van jullie producten?
4. Wat voor impulsen zorgen voor veranderingen die onderhoud aan het model vergen?

Project management

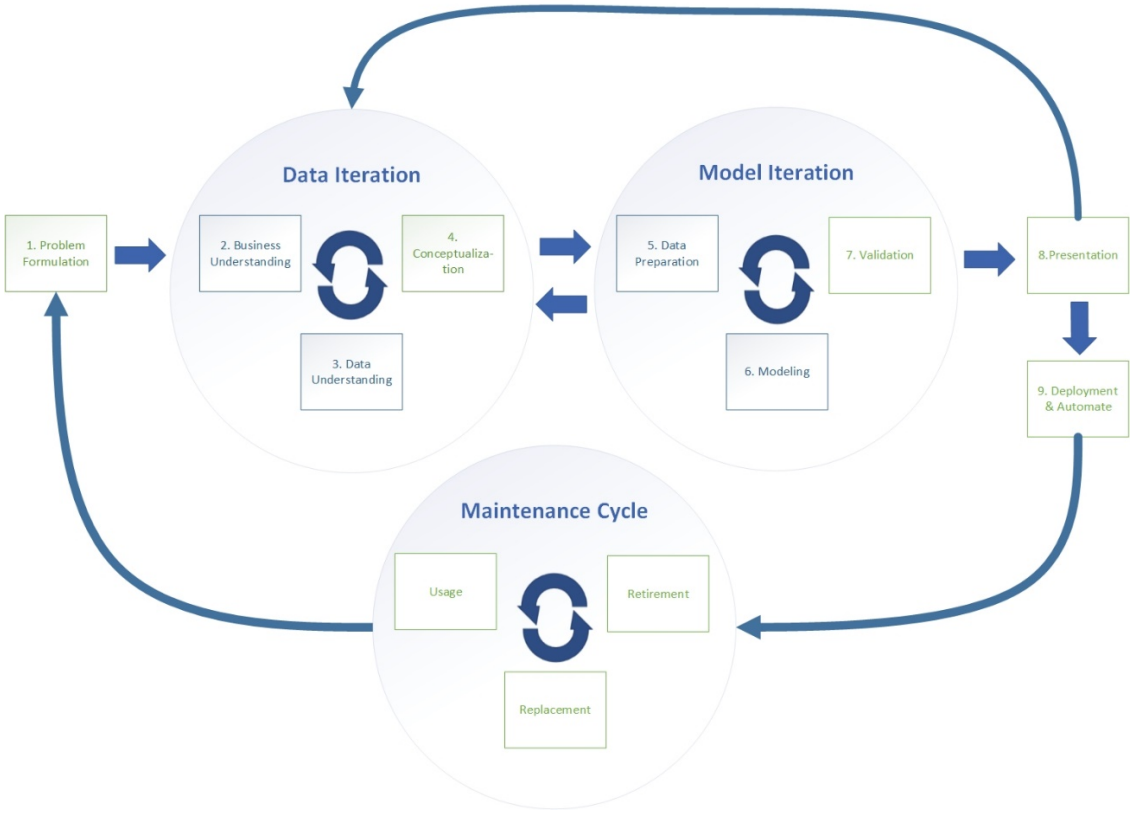
Scrum/Kanban

- Allow capacity-based iterations. Sometimes an iteration takes 1 day, other times 2 weeks. De basis moet een cyclus zijn en niet een hoeveel verbruikte uren.

1. Als je nu kijkt naar de hele cyclus van een project hoe pakken jullie de projecten nu aan?

2. Is dit volledig Agile (volgens het manifest) of is dit waterval of een combinatie van waterval en agile?
3. Waarom werken jullie op deze manier?
4. Welke onderdelen van SCRUM passen jullie toe in de projecten?
5. Kun je hier onderscheid maken in de aanpak tussen de type projecten?
6. Wanneer passen jullie KanBan toe?

Bijlage 4 Focusgroep discussiemodel



Bijlage 5 Codering

List codering aanvangset op basis van het framework en de interviews:

analytics context - hypothese generating
analytics context - hypothese testing
CRISP-DM
data context - big data
data context - data privacy
data context - small data
iterative - Agile-heavy
iterative - SaFe
iterative - hybrid kanban
iterative - hybrid scrum
organisation context - Mandatory Method
organisation context - ROI Culture
project context - problem and innovation oriented
project context - product oriented
team context - multidisciplinair team
team context - open communication
theme - automate
theme - conceptualization
theme - maintenance
theme - Problem formulation